The Annals of Probability
2019, Vol. 47, No. 3, 1270–1334
https://doi.org/10.1214/18-AOP1284

LOCAL SINGLE RING THEOREM ON OPTIMAL SCALE

BY ZHIGANG BAO[1,2], LÁSZLÓ ERDŐS[1] AND KEVIN SCHNELLI[1,3]

HKUST, IST Austria and KTH Royal Institute of Technology

Let U and V be two independent N by N random matrices that are distributed according to Haar measure on $U(N)$. Let Σ be a nonnegative deterministic N by N matrix. The *single ring theorem* [*Ann. of Math. (2)* **174** (2011) 1189–1217] asserts that the empirical eigenvalue distribution of the matrix $X := U\Sigma V^*$ converges weakly, in the limit of large N, to a deterministic measure which is supported on a single ring centered at the origin in $\mathbb{C}$. Within the bulk regime, that is, in the interior of the single ring, we establish the convergence of the empirical eigenvalue distribution on the optimal local scale of order $N^{-1/2+\varepsilon}$ and establish the optimal convergence rate. The same results hold true when U and V are Haar distributed on $O(N)$.

CONTENTS

1. Introduction and main result 1271
   1.1. Local single ring law 1273
   1.2. Summary of previous results 1275
   1.3. Notational conventions 1276
2. Preliminaries and main technical task 1277
   2.1. Free additive convolution 1277
   2.2. The limiting measure $\mu_{\sigma,r}$ 1278
   2.3. Key technical inputs 1280
3. Proof of Theorem 1.8 and Corollary 1.12 1281
4. Local law for block additive model 1285
   4.1. Approximate subordination for block additive models 1288
   4.2. Outline of the strategy of proof 1290
   4.3. Notation 1291
5. Green function subordination for small η 1293
   5.1. Partial randomness decomposition of the Haar measure 1293
   5.2. Green function subordination 1295
   5.3. Recursive moment estimates for $\mathcal{P}_{ii}$ and $\mathcal{K}_{ii}$ 1298
   5.4. Local stability analysis: Proof of Theorem 5.2 1312
   5.5. Continuity argument: Proof of Theorem 5.1 1316
6. Strong law for small η 1321

Received January 2017; revised December 2017.

[1]Supported in part by ERC Advanced Grant RANMAT No. 338804.

[2]Supported in part by Hong Kong RGC Grant ECS 26301517.

[3]Supported in part by the Göran Gustafsson Foundation and the Swedish Research Council Grant VR-2017-05195.

MSC2010 subject classifications. 46L54, 60B20.

Key words and phrases. Non-Hermitian random matrices, local eigenvalue density, single ring theorem, free convolution.

6.1. Proof of Theorem 4.3 on $\mathcal{S}_{\mathcal{I}}(0, \eta_{\mathrm{M}})$ 1322
6.2. Proof of Lemma 6.1 1326
Acknowledgments 1332
Supplementary Material 1332
References 1332

1. Introduction and main result. Consider the $N \times N$ random matrix of the form

$$X \equiv X_N = U \Sigma V^*, \tag{1.1}$$

where $U \equiv U_N$ and $V \equiv V_N$ are two independent sequences of random matrices, which are both Haar distributed on either the unitary group, $U(N)$, of degree N or on the orthogonal group, $O(N)$, of degree N. Moreover, let $\Sigma \equiv \Sigma_N$ be a sequence of $N \times N$ deterministic nonnegative definite diagonal matrices. Note that in general X is not Hermitian and most of its eigenvalues are genuinely complex numbers. In fact, almost surely the matrix X is not normal. Let $\lambda_j(X)$, $j = 1, 2, \ldots, N$, be the eigenvalues of X and let

$$\mu_X := \frac{1}{N} \sum_{j=1}^{N} \delta_{\lambda_j(X)} \tag{1.2}$$

be the (normalized) empirical spectral distribution of X. We define μ_Σ analogously.

ASSUMPTION 1.1. We assume that the sequence (Σ_N) is uniformly bounded, that is, there exists a finite constant S_+ such that

$$0 \le \Sigma_N \le S_+. \tag{1.3}$$

From this assumption, it follows that there is a constant $0 < s_+ < \infty$ such that, for all $N \in \mathbb{N}$,

$$\operatorname{supp} \mu_\Sigma \subset [0, s_+]. \tag{1.4}$$

We first consider the situation where there exists a limiting measure μ_σ[4] of μ_Σ, that is,

$$\mathrm{d_L}(\mu_\Sigma, \mu_\sigma) \to 0, \tag{1.5}$$

as $N \to \infty$, where $\mathrm{d_L}$ denotes the Lévy distance. Given such a μ_σ on $[0, \infty)$, we define

$$r_- := \left(\int_{\mathbb{R}^+} x^{-2} \, \mathrm{d}\mu_\sigma(x) \right)^{-\frac{1}{2}}, \qquad r_+ := \left(\int_{\mathbb{R}^+} x^2 \, \mathrm{d}\mu_\sigma(x) \right)^{\frac{1}{2}}, \tag{1.6}$$

[4]We will often use the convention that capital letters indicate random matrices and the corresponding small letters indicate their limiting objects.

where we set $r_- = 0$ in case the integral in its definition diverges. Note that if μ_σ is supported more than one point, we have $r_- < r_+$ as follows from Schwarz's inequality. We let

$$\mathcal{R}_\sigma \equiv \mathcal{R}(\mu_\sigma) := \{w \in \mathbb{C} : r_- < |w| < r_+\} \tag{1.7}$$

be the ring in $\mathbb{C}$ with radii r_- and r_+. In case $r_- = 0$, $\mathcal{R}_\sigma$ is the punctuated disc of radius r_+.

For a probability measure μ on $\mathbb{R}$, we denote by μ^{sym} its symmetrization, that is, $\mu^{\mathrm{sym}}(A) := \frac{1}{2}[\mu(A) + \mu(-A)]$ for any Borel set $A \subset \mathbb{R}$. For $r \in \mathbb{R}^+$, set

$$\mu_{\sigma,r} := \mu_\sigma^{\mathrm{sym}} \boxplus \delta_r^{\mathrm{sym}}, \tag{1.8}$$

where $\boxplus$ denotes the free additive convolution of probability measures on $\mathbb{R}$; see Section 2.1.

Given a probability measure μ on $\mathbb{R}$, its *Stieltjes' transform*, m_μ, on the complex upper half-plane $\mathbb{C}^+ := \{z \in : \operatorname{Im} z > 0\}$ is defined by

$$m_\mu(z) := \int_{\mathbb{R}} \frac{\mathrm{d}\mu(x)}{x - z}, \qquad z \in \mathbb{C}^+. \tag{1.9}$$

THEOREM 1.2 (Single ring theorem, [26]). *Assume that Assumption 1.1 holds and that there is a compactly supported probability measure μ_σ on $[0, \infty)$, which is supported at more than one point, such that (1.5) holds. Assume in addition that there are constants $k, k_1 > 0$ such that*

$$\operatorname{Im} m_{\mu_\Sigma}(z) \le k_1 \tag{1.10}$$

on $\{z \in \mathbb{C}^+ : \operatorname{Im} z > N^{-k}\}$. Then the empirical spectral distribution μ_X converges weakly (in probability) to a deterministic probability measure ρ_σ supported on $\overline{\mathcal{R}_\sigma}$. The limiting measure is absolutely continuous with respect to Lebesgue measure and given by

$$\rho_\sigma(w)\,\mathrm{d}^2 w = \frac{1}{2\pi}\Delta_w\left(\int_{\mathbb{R}} \log|s|\mu_{\sigma,|w|}(\mathrm{d}s)\right)\mathrm{d}^2 w, \qquad w \in \mathcal{R}_\sigma, \tag{1.11}$$

where $\Delta_w = 4\partial_w\partial_{\overline{w}}$ is the Laplacian on $\mathbb{C}$ and $\mathrm{d}^2 w \equiv \mathrm{d}w \wedge \mathrm{d}\overline{w}$ is Lebesgue measure on $\mathbb{C}$.

REMARK 1.3. In Theorem 1.2, U and V may be both Haar distributed on $U(N)$ or on $O(N)$.

REMARK 1.4. In its original form, Theorem 1.2 was proved by Guionnet, Krishnapur and Zeitouni in [26] under a further assumption on the smallest singular value of the matrix $X - z$, $z \in \mathbb{C}$. This hard-to-check condition was removed by Rudelson and Vershynin in [33] (cf. Theorem 2.6 below), which yields Theorem 1.2.

REMARK 1.5. The measure ρ_σ was first computed in [28]. It has a direct interpretation in free probability theory. In fact, it is the Brown measure of the free product of a Haar unitary and an element σ on a noncommutative probability space; see [28] for more details.

1.1. *Local single ring law.* To state our results, we use the following definition on high-probability estimates from [20]. In Appendix A of the Supplementary Material [5], we collect some of its properties.

DEFINITION 1.6. Let $\mathcal{X} \equiv \mathcal{X}^{(N)}, \mathcal{Y} \equiv \mathcal{Y}^{(N)}$ be two sequences of nonnegative random variables. We say that $\mathcal{Y}$ stochastically dominates $\mathcal{X}$ if, for all (small) $\epsilon > 0$ and (large) $D > 0$,

$$\mathbb{P}\big(\mathcal{X}^{(N)} > N^{\epsilon}\mathcal{Y}^{(N)}\big) \le N^{-D}, \tag{1.12}$$

for sufficiently large $N \ge N_0(\epsilon, D)$, and we write $\mathcal{X} \prec \mathcal{Y}$. When $\mathcal{X}^{(N)}$ and $\mathcal{Y}^{(N)}$ depend on a parameter $v \in \mathcal{V}$ (typically an index label or a spectral parameter), then $\mathcal{X}(v) \prec \mathcal{Y}(v)$, uniformly in $v \in \mathcal{V}$, means that the threshold $N_0(\epsilon, D)$ can be chosen independently of v.

Motivated by (1.11), we introduce a probability measure ρ_Σ on $\mathbb{C}$ by requiring

$$\mathrm{d}\rho_\Sigma(w) = \frac{1}{2\pi}\Delta_w\bigg(\int_{\mathbb{R}} \log|s|\,\mathrm{d}\mu_{\Sigma,|w|}(s)\bigg)\mathrm{d}^2 w, \qquad w \in \mathbb{C}, \tag{1.13}$$

where

$$\mu_{\Sigma,r} := \mu_\Sigma^{\mathrm{sym}} \boxplus \delta_r^{\mathrm{sym}}, \qquad r \ge 0, \tag{1.14}$$

and Δ_w is the Laplacian on $\mathbb{C}$ in the sense of distributions.

REMARK 1.7. The fact that formula (1.13) defines a probability measure follows from previous work on the subject which we shortly summarize here.

Consider a noncommutative W^*-probability space $(\mathcal{M}, \tau)$, with τ a trace. Let u be a Haar unitary element and let $t = t^*$ be $*$-free from u and such that the distribution of t, that is, its spectral measure, is given by μ_Σ. Let $\widetilde{\mu}_{\Sigma,w}$ be the spectral measure of $|ut - w\mathrm{id}|$, with id the unit in $\mathcal{M}$ and $w \in \mathbb{C}$. Then the Brown measure for the product ut is given by the Riesz measure associated to the subharmonic function

$$\mathbb{C} \ni w \mapsto \int_{\mathbb{R}} \log|s|\,\mathrm{d}\widetilde{\mu}_{\Sigma,w}(s); \tag{1.15}$$

cf. Section 2 of [28]. Haagerup and Larsen showed in Proposition 3.5 in [28] that $\widetilde{\mu}_{\Sigma,w} = \mu_{\Sigma,|w|}$. Hence ρ_Σ in (1.13) can be characterized as the Brown measure of ut which by construction is a probability measure.

The main result of this paper is the following local single theorem in the bulk. Notice that (1.5) is not assumed, we only require that $\mathrm{d_L}(\mu_\Sigma, \mu_\sigma) \le b$, for some small constant $b > 0$, for N sufficiently large.

THEOREM 1.8. *Suppose that Assumption 1.1 holds. Let μ_σ be a compactly supported probability measure on $[0, \infty)$ which is supported at more than one point. Fix any (small) $\tau > 0$ and define*

$$\mathcal{R}_\sigma^\tau := \{w \in \mathbb{C} : r_- + \tau \le |w| \le r_+ - \tau\} \subset \mathcal{R}_\sigma, \tag{1.16}$$

where $r_\pm \equiv r_\pm(\mu_\sigma)$ are given in (1.10). Then there exists a (small) constant $b_0 > 0$ and $N_0 \in \mathbb{N}$, depending only on μ_σ and S_+, such that whenever the Lévy distance $\mathrm{d_L}(\mu_\Sigma, \mu_\sigma)$ satisfies

$$\sup_{N \ge N_0} \mathrm{d_L}(\mu_\Sigma, \mu_\sigma) \le b, \tag{1.17}$$

for some $b \le b_0$, then the following holds. Choose any $w_0 \in \mathcal{R}_\sigma^\tau$. Let $f : \mathbb{C} \to \mathbb{R}$ be a smooth function such that $\|f\|_\infty \le C_0$ and $f(z) = 0$ for all $|z| \ge C_0$, for some positive constant C_0. For $\alpha \in (0, 1/2)$, set

$$f_{w_0}(w) := N^{2\alpha} f\big(N^\alpha (w - w_0)\big). \tag{1.18}$$

Then we have for any $\alpha \in (0, 1/2)$ that the estimate

$$\left| \frac{1}{N} \sum_{i=1}^N f_{w_0}(\lambda_i(X)) - \int_{\mathcal{R}_\sigma} f_{w_0}(w)\, \mathrm{d}\rho_\Sigma(w) \right| \prec N^{-1+2\alpha} \|\Delta f\|_{\mathrm{L}^1(\mathbb{C})} \tag{1.19}$$

holds uniformly in f and in $w_0 \in \mathcal{R}_\sigma^\tau$, for N sufficiently large, depending on τ, S_+, μ_σ and C_0.

REMARK 1.9. Note that we can choose α in (1.19), almost as large as $1/2$ in order to have an effective bound on the error term. Since the typical distance between the eigenvalues in the bulk of the ring $\mathcal{R}_\sigma$ is of order $N^{-1/2}$, our result is optimal, both in terms of range of the exponent α and the error term on the right-hand side of (1.19). In particular, this improves the recent local single ring theorem of Benaych–Georges in [11] from scale $(\log N)^{-1/4}$ to the optimal scale $N^{-1/2+\epsilon}$, for any small $\epsilon > 0$.

REMARK 1.10. Theorem 1.8 holds with U, V being Haar distributed on either $U(N)$ or on $O(N)$.

REMARK 1.11. Note that w_0 in Theorem 1.8 is chosen to be the (open) single ring $\mathcal{R}_\sigma$, in particular w_0 stays away from the boundary of $\mathcal{R}_\sigma$. In case $r_- = 0$, $\mathcal{R}_\sigma$ is a punctuated disc. It has been proved in [10, 27] that there are no outliers at an order one distance from $\mathcal{R}_\sigma$.

Let $f : \mathbb{C} \to \mathbb{R}$ be smooth and supported on $\mathcal{R}_\sigma^\tau$, for some (small) $\tau > 0$. Following the proof of Theorem 1.8, it is straightforward to verify that (1.19) also holds with $\alpha = 0$ and f_{w_0} replaced with f, provided that the support of the function f stays away from the spectral edges, that is, is contained in $\mathcal{R}_\sigma^\tau$.

The following corollary of Theorem 1.8 expresses the speed of convergence in the single ring theorem on the macroscopic scale.

COROLLARY 1.12. *Under the conditions and with the notation of Theorem 1.8, we have that*

$$\left| \frac{1}{N} \sum_{i=1}^{N} f\big(\lambda_i(X)\big) - \int_{\mathcal{R}_\sigma} f(w)\, \mathrm{d}\rho_\sigma(w) \right| \prec \|\Delta f\|_{\mathrm{L}^1(\mathbb{C})} \left(\frac{1}{N} + b \right), \tag{1.20}$$

uniformly for any function f supported in $\mathcal{R}_\sigma^\tau$ with a bound $\|f\|_\infty \le C_0$, for N sufficiently large, depending on τ, S_+, μ_σ and C_0.

REMARK 1.13. In (1.20), the measure ρ_σ is given by (1.11). By Theorem 4.4 and Corollary 4.5 of [28], the measure ρ_σ is absolutely continuous on $\mathbb{C} \setminus \{0\}$ with respect to Lebesgue measure. Moreover, it satisfies $\rho_\sigma(\{0\}) = \mu_\sigma(\{0\})$. [In case $\mu_\sigma(\{0\}) > 0$, we have $r_- = 0$.] Note, however, that we have to exclude the point $w = 0$ in our results since it is outside $\mathcal{R}_\sigma$.

REMARK 1.14. Note that in Theorem 1.8 and Corollary 1.12 we do not require any regularity assumption on the measure μ_σ; we even allow for atoms in μ_σ. In particular, sending $b \to 0$, as $N \to \infty$, Corollary 1.12 also implies that Assumption 1.1 and (1.5) together imply $\mathrm{d}_{\mathrm{L}}(\rho_\Sigma, \rho_\sigma) \to 0$, as $N \to \infty$, thus removing the regularity condition (1.10) in the bulk from the single ring theorem, this answers a question in [26], Remark 2.

1.2. *Summary of previous results.* The first single ring theorem was established by Feinberg and Zee for a class of unitary invariant ensemble in [24], but without full rigor. The complete mathematical proof was given by Guionnet, Krishnapur and Zeitouni [26]; see also Remarks 1.4 and 1.14 for relaxing some conditions.

In spirit of the Wigner ensemble for the Hermitian case, the Ginibre ensemble can also be naturally extended by considering arbitrary i.i.d. entries; however, the unitary invariance property is lost in this generalization. Starting from the work of Girko [25], until the final result of Tao and Vu [34] with the least moment assumption, there have been many works devoted in proving circular law for general distribution. We refer to the survey [14] for more references in this direction. A prominent idea called *Hermitization* was introduced by Girko in [25]. This method translates spectral distribution problems of a non-Hermitian matrix to

those of a Hermitian matrix (of double dimension), whose spectral properties can be studied with more established techniques.

Similar to Wigner's original semicircle law, the single ring theorem establishes weak convergence of the spectral distribution, that is, it captures the density of eigenvalues on the global scale. Since the typical distance between nearby eigenvalues is very small, of order $N^{-1/2}$, it is natural to ask whether the empirical density can also be approximated by the deterministic limit density on some local scale. Ideally, such *local law* should hold on the smallest possible scale, that is, just above the scale $N^{-1/2}$. In the Hermitian case, the local laws for Wigner and related ensembles have been extensively studied in the recent years (see, e.g. [19] for a survey and references therein), the optimal local scale has been first achieved in [21].

With the aid of Girko's Hermitization, local laws for non-Hermitian matrices can be obtained via studying the local law for certain Hermitian matrices. With this strategy, the local circular law on optimal scale was established in the series of works Bourgade, Yau and Yin [15, 16] and Yin [37]. The first local single ring theorem was obtained by Benaych–Georges in [11], down to the scale $(\log N)^{-\frac{1}{4}}$, by proving the matrix subordination for Girko's Hermitization of X in (1.1); cf. (2.14). The strategy of matrix subordination was originally introduced by Kargin in [29] for proving a local law in the additive matrix model $A + UBU^*$, where A and B are deterministic Hermitian matrices and U is a Haar unitary. This additive model shares certain similarities with the Hermitization of the model $X = U\Sigma V^*$, but the latter has a block structure, and thus we call it *block additive model* (cf. (4.2)). Recently, in [3, 4, 6], we obtained the local law of the additive model $A + UBU^*$ on the optimal scale. The approach developed in these works opens up a path to treat the optimal local law in the block additive model, hence also sheds light on the optimal local single ring theorem. The key difference is that in the block additive model the Haar unitary matrices provide only a randomized $U(N) \times U(N)$ symmetry instead of the full $U(2N)$ symmetry. In particular, the coupling between the blocks is deterministic, so the mixing mechanism is much weaker. A more detailed overview of the proof strategy and the difficulties will be given in Section 4.2.

1.3. *Notational conventions*. We use the symbols $O(\cdot)$ and $o(\cdot)$ for the standard big-O and little-o notation. We use c and C to denote strictly positive constants that do not depend on N. Their values may change from line to line.

We denote by $M_N(\mathbb{C})$ the set of $N \times N$ matrices over $\mathbb{C}$. For $A \in M_N(\mathbb{C})$, we denote by $\|A\|$ its operator norm and by $\|A\|_2$ its Hilbert–Schmidt norm. The matrix entries of A are denoted by A_{ij}.

Let $\boldsymbol{g} = (g_1, \dots, g_N)$ be a real or complex Gaussian vector. We write $\boldsymbol{g} \sim \mathcal{N}_{\mathbb{R}}(0, \sigma^2 I_N)$ if $g_1, \dots, g_N$ are independent and identically distributed (i.i.d.) $N(0, \sigma^2)$ normal variables; and we write $\boldsymbol{g} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2 I_N)$ if $g_1, \dots, g_N$ are i.i.d.

$N_{\mathbb{C}}(0,\sigma^2)$ variables, where $g_i \sim N_{\mathbb{C}}(0,\sigma^2)$ means that $\operatorname{Re} g_i$ and $\operatorname{Im} g_i$ are independent $N(0,\frac{\sigma^2}{2})$ normal variables.

We use double brackets to denote index sets, that is, for $n_1, n_2 \in \mathbb{R}$, $[\![n_1, n_2]\!] := [n_1, n_2] \cap \mathbb{Z}$.

2. Preliminaries and main technical task.

2.1. *Free additive convolution.* We recall some basic notions and results for the free additive convolution. We follow the notational conventions in our previous paper [2].

Let μ be a Borel probability measure on $\mathbb{R}$ and recall its Stieltjes' transform m_μ defined in (1.9). Note that $m_\mu : \mathbb{C}^+ \to \mathbb{C}^+$ is an analytic function such that

$$\lim_{\eta \nearrow \infty} \mathrm{i}\eta m_\mu(\mathrm{i}\eta) = -1. \tag{2.1}$$

Conversely, if $m : \mathbb{C}^+ \to \mathbb{C}^+$ is an analytic function such that $\lim_{\eta\nearrow\infty} \mathrm{i}\eta m(\mathrm{i}\eta) = -1$, then m is the Stieltjes' transform of a probability measure μ.

Given a Borel probability measure μ on $\mathbb{R}$, let F_μ be the *negative reciprocal Stieltjes transform* of μ,

$$F_\mu(z) := -\frac{1}{m_\mu(z)}, \qquad z \in \mathbb{C}^+. \tag{2.2}$$

Observe that

$$\lim_{\eta \nearrow \infty} \frac{F_\mu(\mathrm{i}\eta)}{\mathrm{i}\eta} = 1, \tag{2.3}$$

as follows from (2.1). Note that F_μ is analytic on $\mathbb{C}^+$ with nonnegative imaginary part.

The *free additive convolution* is the symmetric binary operation on Borel probability measures on $\mathbb{R}$ characterized by the following result.

THEOREM 2.1 (Theorem 4.1 in [9], Theorem 2.1 in [17]). *Given two Borel probability measures, μ_1 and μ_2, on $\mathbb{R}$, there exist unique analytic functions, $\omega_1, \omega_2 : \mathbb{C}^+ \to \mathbb{C}^+$, such that*:

(i) *for all $z \in \mathbb{C}^+$, $\operatorname{Im}\omega_1(z), \operatorname{Im}\omega_2(z) \geq \operatorname{Im} z$ and*

$$\lim_{\eta \nearrow \infty} \frac{\omega_1(\mathrm{i}\eta)}{\mathrm{i}\eta} = \lim_{\eta \nearrow \infty} \frac{\omega_2(\mathrm{i}\eta)}{\mathrm{i}\eta} = 1; \tag{2.4}$$

(ii) *for all $z \in \mathbb{C}^+$,*

$$F_{\mu_1}\big(\omega_2(z)\big) = F_{\mu_2}\big(\omega_1(z)\big), \qquad \omega_1(z) + \omega_2(z) - z = F_{\mu_1}\big(\omega_2(z)\big). \tag{2.5}$$

It follows from (2.4) that the analytic function $F : \mathbb{C}^+ \to \mathbb{C}^+$ defined by

$$F(z) := F_{\mu_1}(\omega_2(z)) = F_{\mu_2}(\omega_1(z)), \tag{2.6}$$

satisfies the analogue of (2.3). Thus F is the negative reciprocal Stieltjes' transform of a probability measure μ, called the free additive convolution of μ_1 and μ_2, denoted by $\mu \equiv \mu_1 \boxplus \mu_2$. The functions ω_1 and ω_2 are referred to as the *subordination functions* and F is said to be subordinated to F_{μ_1}, respectively, to F_{μ_2}. The subordination phenomenon was first noted by Voiculescu [36] in a generic situation and extended to full generality by Biane [13]. To exclude trivial shifts of measures, we henceforth assume that both, μ_1 and μ_2, are supported at more than one point. Then the analytic functions F, ω_1 and ω_2 extend continuously to the real line; see Theorem 2.3 [7] or Theorem 3.3 [8]. We use the same notation for their extensions to $\mathbb{C}^+ \cup \mathbb{R}$.

2.2. *The limiting measure $\mu_{\sigma,r}$.* Recall the definitions $\mu_{\Sigma,r} := \mu_\Sigma^{\mathrm{sym}} \boxplus \delta_r^{\mathrm{sym}}$ and $\mu_{\sigma,r} := \mu_\sigma^{\mathrm{sym}} \boxplus \delta_r^{\mathrm{sym}}$ from (1.8). In this subsection, we will always assume that μ_Σ and μ_σ satisfy Assumption 1.1. For the sake of simplicity of notation, we abbreviate in this subsection

$$\mu_1 \equiv \mu_\sigma^{\mathrm{sym}}, \qquad \mu_2 \equiv \delta_r^{\mathrm{sym}}. \tag{2.7}$$

The negative reciprocal Stieltjes' transform of $\mu_2 = \delta_r^{\mathrm{sym}}$ is found to be

$$F_{\mu_2}(z) = z - \frac{r^2}{z}, \qquad z \in \mathbb{C}^+. \tag{2.8}$$

Substituting (2.8) into (2.5), we obtain

$$F_{\mu_1}(\omega_2(z)) = F_{\mu_2}(\omega_1(z)) = F_{\mu_1}(\omega_2(z)) - \omega_2(z) + z - \frac{r^2}{F_{\mu_1}(\omega_2(z)) - \omega_2(z) + z}.$$

Solving the above equation for $F_{\mu_1}(\omega_2(z))$ we conclude that the subordination function $\omega_2(z)$ is the unique solution to

$$F_{\mu_1}(\omega_2(z)) - \omega_2(z) = -z - \frac{r^2}{\omega_2(z) - z}, \qquad z \in \mathbb{C}^+, \tag{2.9}$$

subject to the condition $\operatorname{Im}\omega_2(z) \geq \operatorname{Im} z$. Comparing once more with (2.5) we immediately find that the other subordination function is given by

$$\omega_1(z) = -\frac{r^2}{\omega_2(z) - z}, \qquad z \in \mathbb{C}^+. \tag{2.10}$$

The analysis of the measure $\mu_{\sigma,r} = \mu_1 \boxplus \mu_2$ thus reduces to the analysis of (2.9) for ω_2. We first derive upper and lower bound on $\omega_2(z)$. For the purpose of proving Theorem 1.8, it will suffice to consider $z \in \{\mathrm{i}\eta : \eta \geq 0\}$. Since μ_1 and μ_2 are symmetric, we have $\omega_2(\mathrm{i}\eta) = -\overline{\omega_2(\mathrm{i}\eta)}$, that is, $\omega_2(\mathrm{i}\eta)$ and $\omega_1(\mathrm{i}\eta)$ are both

fully imaginary. This simplifies our analysis; while detailed quantitative properties of the full measure $\mu_{\sigma,r}$ are still poorly understood, we now have a good control on it near zero, hence on its Stieltjes' transform along the imaginary axis. The main result, formulated in Theorem 2.2 below, is that the subordination functions are bounded from below and above on the imaginary axis without any condition on μ_σ. This theorem is the key input that enables us to dispense with the regularity condition in the single ring theorem; see Remark 1.14.

THEOREM 2.2 (Bounds on subordination functions). *We assume that the support of μ_σ contains more than one point, equivalently, that $r_- < r_+$. Let $\mu_1 = \mu_\sigma^{\mathrm{sym}}$ and $\mu_2 = \delta_r^{\mathrm{sym}}$ for some $r > 0$. Fix $\eta_{\mathrm{M}} < \infty$ and a (small) $\tau > 0$. Set*

$$J := [r_- + \tau, r_+ - \tau].$$

There exist constants $c \equiv c(\mu_1, \tau, \eta_{\mathrm{M}}) > 0$ and $C \equiv C(\mu_1, \tau, \eta_{\mathrm{M}}) < \infty$ such that

$$\sup_{r\in J}\sup_{\eta\in[0,\eta_{\mathrm{M}}]}\big|\omega_1(\mathrm{i}\eta)\big| \le C, \qquad \sup_{r\in J}\sup_{\eta\in[0,\eta_{\mathrm{M}}]}\big|\omega_2(\mathrm{i}\eta)\big| \le C, \tag{2.11}$$

$$\inf_{r\in J}\inf_{\eta\in[0,\eta_{\mathrm{M}}]}\operatorname{Im}\omega_1(\mathrm{i}\eta) \ge c, \qquad \inf_{r\in J}\inf_{\eta\in[0,\eta_{\mathrm{M}}]}\operatorname{Im}\omega_2(\mathrm{i}\eta) \ge c \tag{2.12}$$

and

$$\inf_{r\in J}\inf_{\eta\in[0,\eta_{\mathrm{M}}]}\big|m_{\mu_1\boxplus\mu_2}(\mathrm{i}\eta)\big| \ge c, \sup_{r\in J}\sup_{\eta\in[0,\eta_{\mathrm{M}}]}\big|m_{\mu_1\boxplus\mu_2}(\mathrm{i}\eta)\big| \le C. \tag{2.13}$$

REMARK 2.3. By (2.13), the measure $\mu_1 \boxplus \mu_2$ has a positive and bounded density at $E = 0$. In particular, $E = 0$ is in the bulk of the measure $\mu_1 \boxplus \mu_2$, as defined in Definition 4.2 below.

The proof of Theorem 2.2 is quite technical and independent of the main line of the argument, so we give it in Section 7 of the Supplementary Material [5]. In the subsequent sections, we will mainly rely on the following corollary of Theorem 2.2. Let $m_{\Sigma,r}(z)$ be the Stieltjes' transform of $\mu_{\Sigma,r}$; see (1.8).

COROLLARY 2.4. *Fix $\eta_{\mathrm{M}} < \infty$ and a (small) $\tau > 0$. Then there are constants $C \equiv C(\mu_\sigma^{\mathrm{sym}}, \tau, \eta_{\mathrm{M}})$, $c \equiv (\mu_\sigma^{\mathrm{sym}}, \tau, \eta_{\mathrm{M}})$ and a threshold $N_0 \equiv N_0(\mu_\sigma^{\mathrm{sym}}, \tau, \eta_{\mathrm{M}})$ such that the conclusions in Theorem 2.2 hold with $\mu_1 = \mu_\Sigma^{\mathrm{sym}}$ and $\mu_2 = \delta_r^{\mathrm{sym}}$, for $N \ge N_0$.*

PROOF. This follows directly from the continuity of the subordination functions with respect to the Lévy distance (see Lemma 5.1 of [2]), from Theorem 2.2 and from (1.17). □

2.3. *Key technical inputs*. Following Girko's hermitization technique [25], we introduce for any $w \in \mathbb{C}$ the $2N \times 2N$ Hermitian matrix

$$H^w := \begin{pmatrix} 0 & X - w \\ X^* - w^* & 0 \end{pmatrix}. \tag{2.14}$$

The main advantage of working with H^w is that it is self-adjoint and we thus have a functional calculus at disposal. For any function $g \in C^2(\mathbb{C})$, an application of Green's theorem reveals that

$$\frac{1}{N}\sum_{i=1}^{N} g\big(\lambda_i(X)\big) = \frac{1}{2\pi}\int_{\mathbb{C}} (\Delta g)(w)\left(\frac{1}{2N}\operatorname{Tr}\log\big|H^w\big|\right)\mathrm{d}^2 w, \tag{2.15}$$

which is a manifestation of $\log|\cdot|$ being the Coulomb potential in two dimensions. The following identity, first used in this context by [35], allows us to efficiently deal with the right-hand side of (2.15). For any (large) $K > 0$,

$$\frac{1}{2N}\operatorname{Tr}\log\big|H^w\big| = \frac{1}{2N}\operatorname{Tr}\log\big|(H^w - \mathrm{i}K)\big| - \operatorname{Im}\int_0^K m^w(\mathrm{i}\eta)\,\mathrm{d}\eta, \tag{2.16}$$

with $|w| > 0$, where $m^w(z)$, $z \in \mathbb{C}^+$, is the Stieltjes' transform of the spectral distribution of H^w. For very large K the first term on the right-hand side of (2.16) is elementary to control, we hence focus on the second term. Due to the block structure of H^w, the eigenvalues come in pairs $\pm\lambda_i^w$, $i \in [\![1, N]\!]$, where $0 \le \lambda_1^w \le \cdots \le \lambda_N^w$ are the nonnegative eigenvalues. With this notation, m^w is given by

$$m^w(z) := \frac{1}{2N}\sum_{i=1}^{N}\left(\frac{1}{\lambda_i^w - z} + \frac{1}{-\lambda_i^w - z}\right) = \frac{1}{N}\sum_{i=1}^{N}\frac{\lambda_i^w}{(\lambda_i^w)^2 - z^2}, \qquad z \in \mathbb{C}^+.$$

Recall the notation $m_{\Sigma,|w|}$ for the Stieltjes' transform of $\mu_{\Sigma,|w|}$; cf. (1.8). The following result is the main technical input for the proof of Theorem 1.8. Recall $\mathcal{R}_\sigma^\tau$ from (1.16).

THEOREM 2.5 (Local law for H^w). *Under the conditions and with the notations of Theorem* 1.8, *the estimate*

$$\sup_{w\in\mathcal{R}_\sigma^\tau}\big|m^w(\mathrm{i}\eta) - m_{\Sigma,|w|}(\mathrm{i}\eta)\big| \prec \frac{1}{N\eta}, \tag{2.17}$$

holds uniformly in $\eta > 0$, *for* N *sufficiently large*, *depending on* τ, S_+ *and* μ_σ.

This result controls $|m^w(\mathrm{i}\eta) - m_{\Sigma,|w|}(\mathrm{i}\eta)|$ along the positive imaginary axis. Note that the error estimate on the right-hand side of (2.17) is effective when η is chosen just above the local scale, that is, when $\eta > N^{-1+\gamma}$, for any small $\gamma > 0$. For even smaller $\eta > 0$, (2.17) yields the upper bound $|m^w(\mathrm{i}\eta)| \prec (N\eta)^{-1}$ which improves the trivial deterministic bound $|m^w(\mathrm{i}\eta)| \le \eta^{-1}$ by a factor N^{-1}. Theorem 2.5 is used to control the integrand in the second term on the right-hand side

of (2.16) for $\eta \gtrsim N^{-1}$. On very short scales, the behavior of $m^w(\mathrm{i}\eta)$, $\eta \lesssim N^{-1}$, is essentially random and determined by the smallest (in absolute value) eigenvalues of H^w. The following estimate on λ_1^w, proved by Rudelson and Vershynin in [33], is then used to control the integrand of the second term on the right-hand side of (2.16) for very small $\eta \lesssim N^{-1}$.

THEOREM 2.6 (Theorem 1.1 and Theorem 1.2 in [33]). *There exist positive numerical constants $c > 0$ and $C < \infty$, such that*

$$\mathbb{P}\left(\lambda_1^w \le \frac{t}{|w|}\right) \le \left(\frac{t}{|w|}\right)^c N^C, \tag{2.18}$$

uniformly in $t > 0$, for all $N \in \mathbb{N}$.

REMARK 2.7. In the orthogonal case, (2.18) holds, for N sufficiently large, when the matrix Σ is away from the identity; see Theorem 1.2 in [33]. In this case, the constants c, C and the threshold for N in (2.18) depend on S_+ and μ_σ. Indeed, (1.17) and the assumption that the support of μ_σ contains more than one point imply that Σ is separated away from the identity.

In Section 3, we will choose g in (2.15) to be the rescaled function $f_{w_0}(\cdot) = N^{2\alpha} f(N^\alpha(\cdot - w_0))$; see (1.18). The local law in (2.17) together with (2.18) (with $t/|w| \ll N^{-1}$) will allow us to choose $\alpha \in (0, 1/2)$ as is asserted in Theorem 1.8. The details of the proof of Theorem 1.8, assuming Theorem 2.5, are carried out in Section 3. Our main task then is to prove Theorem 2.5. Actually, we will establish the local law in a more general setting; cf. Theorem 4.3. This will be accomplished in Sections 4–6 and we will separately outline the main ideas of this proof in Section 4.2. We begin with the proof of Theorem 2.2 in the next section.

3. Proof of Theorem 1.8 and Corollary 1.12. In this section we prove Theorem 1.8 and Corollary 1.12, with the aid of Theorems 2.5 and 2.6. The use of Girko's hermitized matrices to derive local laws is a standard argument; see, for example, [15, 35] for related models. Following [35], we use the identity (3.6) below to link the log-determinant of H^w with the Stieltjes' transform m^w.

PROOF OF THEOREM 1.8. For any $\zeta \in \mathbb{C}$, we denote

$$w \equiv w(\zeta) := w_0 + N^{-\alpha}\zeta. \tag{3.1}$$

Given $f : \mathbb{C} \to \mathbb{R}$ satisfying the assumption of Theorem 1.8, we introduce the domain

$$\mathcal{D}_{w_0}(\alpha) \equiv \mathcal{D}_{w_0}(\alpha, f) := \{\tilde{w} : N^\alpha(\tilde{w} - w_0) \in \mathrm{supp}(f)\}. \tag{3.2}$$

According to (3.1), $w \in \mathcal{D}_{w_0}(\alpha)$ is equivalent to $\zeta \in \operatorname{supp}(f)$, in particular $|\zeta| \le C$ as f is compactly supported. Recall the notation $f_{w_0}(\cdot)$ from Theorem 1.8. Using (2.15), we rewrite

$$\frac{1}{N}\sum_i f_{w_0}(\lambda_i(X)) = \frac{1}{2\pi}N^{2\alpha}\int_{\mathbb{C}}(\Delta f)(\zeta)\left(\frac{1}{2N}\operatorname{Tr}\log|H^w|\right)\mathrm{d}^2\zeta. \tag{3.3}$$

Recalling the definitions in (1.8) and (1.13), we also have

$$\begin{aligned}\int_{\mathbb{C}} f_{w_0}(w)\rho_\Sigma(\mathrm{d}^2 w) &= \frac{1}{2\pi}\int_{\mathbb{C}} f_{w_0}(w)\Delta_w\left(\int_{\mathbb{R}}\log|u|\mu_{\Sigma,|w|}(\mathrm{d}u)\right)\mathrm{d}^2 w \\ &= \frac{1}{2\pi}N^{2\alpha}\int_{\mathbb{C}}(\Delta f)(\zeta)\left(\int_{\mathbb{R}}\log|u|\mu_{\Sigma,|w|}(\mathrm{d}u)\right)\mathrm{d}^2\zeta.\end{aligned} \tag{3.4}$$

Hence, we can write

$$\begin{aligned}&\frac{1}{N}\sum_i f_{w_0}(\lambda_i(X)) - \int_{\mathbb{C}} f_{w_0}(w)\rho_\Sigma(\mathrm{d}^2 w) \\ &\quad= \frac{1}{2\pi}N^{2\alpha}\int_{\mathbb{C}}(\Delta f)(\zeta)\left(\frac{1}{2N}\operatorname{Tr}\log|H^w| - \int_{\mathbb{R}}\log|u|\mu_{\Sigma,|w|}(\mathrm{d}u)\right)\mathrm{d}^2\zeta.\end{aligned} \tag{3.5}$$

We next use the following observation due to [35], Section 8. For any (large) $K > 0$ and $|w| > 0$, we have

$$\frac{1}{2N}\operatorname{Tr}\log|H^w| = \frac{1}{2N}\operatorname{Tr}\log|(H^w - \mathrm{i}K)| - \operatorname{Im}\int_0^K m^w(\mathrm{i}\eta)\,\mathrm{d}\eta. \tag{3.6}$$

Analogously, we can also write, with the same K,

$$\int_{\mathbb{R}}\log|u|\mu_{\Sigma,|w|}(\mathrm{d}u) = \int_{\mathbb{R}}\log|u-\mathrm{i}K|\mu_{\Sigma,|w|}(\mathrm{d}u) - \operatorname{Im}\int_0^K m_{\Sigma,|w|}(\mathrm{i}\eta)\,\mathrm{d}\eta. \tag{3.7}$$

Choosing K sufficiently large, say $K = N^L$ for some large constant L, it is easy to see that

$$\left|\frac{1}{2N}\operatorname{Tr}\log|(H^w - \mathrm{i}K)| - \int_{\mathbb{R}}\log|u-\mathrm{i}K|\mu_{\Sigma,|w|}(\mathrm{d}u)\right| \ll \frac{1}{N} \tag{3.8}$$

holds uniformly in $w \in \mathcal{D}_{w_0}(\alpha)$. Here, we used the fact that $\|H^w\| \le C$ for some positive constant C; cf. Assumption 1.1. The uniformity in w can be guaranteed by the fact that $\mathcal{D}_{w_0}(\alpha)$ lies in a ball of finite (in fact, $CN^{-\alpha}$) radius since f is compactly supported. Hence, it suffices to show

$$\left|\int_{\mathbb{C}}(\Delta f)(\zeta)\left(\operatorname{Im}\int_0^{N^L}\left(m^w(\mathrm{i}\eta) - m_{\Sigma,|w|}(\mathrm{i}\eta)\right)\mathrm{d}\eta\right)\mathrm{d}^2\zeta\right| \prec \frac{\|\Delta f\|_{\mathrm{L}^1(\mathbb{C})}}{N}. \tag{3.9}$$

To show (3.9), we decompose the integral with respect to η into two parts:

$$\int_0^{N^L} = \int_0^{N^{-L_1}} + \int_{N^{-L_1}}^{N^L}, \tag{3.10}$$

for sufficiently large constants $L_1 > 1$ and $L > 0$ to be chosen below. To control the first part, we use (2.18), while for the second part we use (2.17).

First, using the upper bound of $m_{\Sigma,|w|}(\mathrm{i}\eta)$ (cf. Corollary 2.4), we obtain

$$\left|\int_0^{N^{-L_1}} \operatorname{Im} m_{\Sigma,|w|}(\mathrm{i}\eta)\,\mathrm{d}\eta\right| \le \frac{1}{N}, \tag{3.11}$$

for $L_1 > 1$, uniformly in $w \in \mathcal{D}_{w_0}(\alpha)$. Hence, we have

$$\left|\int_{\mathbb{C}} (\Delta f)(\zeta)\left(\int_0^{N^{-L_1}} \operatorname{Im} m_{\Sigma,|w|}(\mathrm{i}\eta)\,\mathrm{d}\eta\right)\mathrm{d}^2\zeta\right| \le C\frac{\|\Delta f\|_{\mathrm{L}^1(\mathbb{C})}}{N}. \tag{3.12}$$

In addition, we observe that

$$\begin{aligned}
&\mathbb{P}\left(\left|\int_{\mathbb{C}} (\Delta f)(\zeta)\left(\int_0^{N^{-L_1}} \operatorname{Im} m^w(\mathrm{i}\eta)\,\mathrm{d}\eta\right)\mathrm{d}^2\zeta\right| > \frac{\|\Delta f\|_{\mathrm{L}^1(\mathbb{C})}}{N}\right)\\
&\quad\le \frac{N}{\|\Delta f\|_{\mathrm{L}^1(\mathbb{C})}}\mathbb{E}\left|\int_{\mathbb{C}} (\Delta f)(\zeta)\left(\int_0^{N^{-L_1}} \operatorname{Im} m^w(\mathrm{i}\eta)\,\mathrm{d}\eta\right)\mathrm{d}^2\zeta\right| \qquad (3.13)\\
&\quad\le \frac{N}{\|\Delta f\|_{\mathrm{L}^1(\mathbb{C})}}\int_{\mathbb{C}} |(\Delta f)(\zeta)|\mathbb{E}\left(\int_0^{N^{-L_1}} \frac{\eta}{(\lambda_1^w)^2+\eta^2}\,\mathrm{d}\eta\right)\mathrm{d}^2\zeta.
\end{aligned}$$

Note that

$$\begin{aligned}
&\mathbb{E}\left(\int_0^{N^{-L_1}} \frac{\eta}{(\lambda_1^w)^2+\eta^2}\,\mathrm{d}\eta\right)\\
&\quad= \frac{1}{2}\mathbb{E}\log\bigl(1+\bigl(N^{L_1}\lambda_1^w\bigr)^{-2}\bigr)\\
&\quad= \frac{1}{2}\int_0^{\infty} \mathbb{P}\bigl(\log\bigl(1+\bigl(N^{L_1}\lambda_1^w\bigr)^{-2}\bigr)\ge s\bigr)\,\mathrm{d}s\\
&\quad= \frac{1}{2}\int_0^{\infty} \mathbb{P}\bigl(\lambda_1^w \le N^{-L_1}\bigl(\mathrm{e}^s-1\bigr)^{-\frac{1}{2}}\bigr)\,\mathrm{d}s\\
&\quad= \frac{1}{2}\left(\int_0^{N^{-L_1}} + \int_{N^{-L_1}}^1 + \int_1^{\infty}\right)\mathbb{P}\bigl(\lambda_1^w \le N^{-L_1}\bigl(\mathrm{e}^s-1\bigr)^{-\frac{1}{2}}\bigr)\,\mathrm{d}s.
\end{aligned}$$

For the first integral, we use the trivial bound $\mathbb{P}(\cdot) \le 1$ to obtain

$$\int_0^{N^{-L_1}} \mathbb{P}\bigl(\lambda_1^w \le N^{-L_1}\bigl(\mathrm{e}^s-1\bigr)^{-\frac{1}{2}}\bigr)\,\mathrm{d}s \le N^{-L_1}. \tag{3.14}$$

For the second part of the integral, using the crude bound $(\mathrm{e}^s-1)^{-\frac{1}{2}} \le s^{-\frac{1}{2}} \le N^{\frac{L_1}{2}}$, $s \in [N^{-L_1}, 1]$ and (2.18), we estimate

$$\int_{N^{-L_1}}^1 \mathbb{P}\bigl(\lambda_1^w \le N^{-L_1}\bigl(\mathrm{e}^s-1\bigr)^{-\frac{1}{2}}\bigr)\,\mathrm{d}s \le \int_{N^{-L_1}}^1 \mathbb{P}\bigl(\lambda_1^w \le N^{-\frac{L_1}{2}}\bigr)\,\mathrm{d}s \le N^{-\frac{cL_1}{2}+C},$$

for some constants $c > 0$ and $C < \infty$, for N sufficiently large. For the third part, using $\mathrm{e}^s - 1 > \frac{1}{2}\mathrm{e}^s$, $s > 1$, and (2.18), we have

$$\begin{aligned}\int_1^\infty \mathbb{P}\big(\lambda_1^w \le N^{-L_1}(\mathrm{e}^s-1)^{-\frac{1}{2}}\big)\,\mathrm{d}s &\le \int_1^\infty \mathbb{P}\big(\lambda_1^w \le \sqrt{2}N^{-L_1}\mathrm{e}^{-\frac{s}{2}}\big)\,\mathrm{d}s \\ &\le \frac{N^{-cL_1+C}}{2}\int_1^\infty \mathrm{e}^{-\frac{cs}{2}}\,\mathrm{d}s \le N^{-cL_1+C},\end{aligned} \tag{3.15}$$

for some constants $c > 0$ and $C < \infty$. Combining (3.14)–(3.15), we obtain that there are positive constants $c' > 0$ and C', independent of L_1 such that

$$\mathbb{E}\bigg(\int_0^{N^{-L_1}} \frac{\eta}{(\lambda_1^w)^2+\eta^2}\,\mathrm{d}\eta\bigg) \le N^{-c'L_1+C'}, \tag{3.16}$$

for N sufficiently large. In fact, the bound (3.16) is uniform in $w \in \mathcal{D}_{w_0}(\alpha)$ since the constants c and C in Theorem 2.6 are uniform in t and w. Plugging (3.16) into (3.13), yields

$$\begin{aligned}&\mathbb{P}\bigg(\bigg|\int_{\mathbb{C}}(\Delta f)(\zeta)\bigg(\int_0^{N^{-L_1}} \operatorname{Im} m^w(\mathrm{i}\eta)\,\mathrm{d}\eta\bigg)\mathrm{d}^2\zeta\bigg| \ge \frac{\|\Delta f\|_{\mathrm{L}^1(\mathbb{C})}}{N}\bigg) \\ &\quad\le N^{-c'L_1+C'+1},\end{aligned} \tag{3.17}$$

for N sufficiently large (independent of L_1). Choosing L_1 large enough, the contribution of the first integral in (3.10) to (3.9) is within the claimed error.

To control the contributions from the second integral in (3.10), for any (large) constant L_1, we apply the local law for m^w in (2.17), uniform in w, to find

$$\begin{aligned}&\bigg|\int_{\mathbb{C}}(\Delta f)(\zeta)\bigg(\operatorname{Im}\int_{N^{-L_1}}^{N^L}\big(m^w(\mathrm{i}\eta)-m_{\Sigma,|w|}(\mathrm{i}\eta)\big)\,\mathrm{d}\eta\bigg)\mathrm{d}^2\zeta\bigg| \\ &\quad\prec \int_{\mathbb{C}}\big|(\Delta f)(\zeta)\big|\bigg(\int_{N^{-L_1}}^{N^L}\frac{1}{N\eta}\,\mathrm{d}\eta\bigg)\mathrm{d}^2\zeta \prec \frac{\|\Delta f\|_{\mathrm{L}^1(\mathbb{C})}}{N}.\end{aligned}$$

Combining (3.12) and (3.17), and choosing L_1 sufficiently large, we get (3.9), which together with (3.5)–(3.8) concludes the proof of Theorem 1.8. □

PROOF OF COROLLARY 1.12. Let $f : \mathbb{C} \to \mathbb{R}$ be smooth and supported on $\mathcal{R}_\sigma^\tau$; see (1.16). It is straightforward following the proof of Theorem 1.8 to verify that (1.19) also holds with $\alpha = 0$ and f_{w_0} replaced with f provided that $\operatorname{supp} f \subset \mathcal{R}_\sigma^\tau$; cf. Remark 1.11. Thus under the assumptions of Corollary 1.12 it suffices to show that

$$\begin{aligned}&\bigg|\int_{\mathcal{R}_\sigma^\tau}(\Delta f)(w)\int_{\mathbb{R}}\log|u|\big(\mu_{\Sigma,|w|}(\mathrm{d}u)-\mu_{\sigma,|w|}(\mathrm{d}u)\big)\,\mathrm{d}^2w\bigg| \\ &\quad\le C\|\Delta f\|_{\mathrm{L}^1(\mathbb{C})}\,\mathrm{d}_{\mathrm{L}}(\mu_\Sigma,\mu_\sigma),\end{aligned}$$

for a constant C (depending on τ), to conclude its proof. From (3.7), it is sufficient to prove that

$$\left|\int_{\mathbb{R}} \log|u-\mathrm{i}|\big(\mu_{\Sigma,|w|}(\mathrm{d}u)-\mu_{\sigma,|w|}(\mathrm{d}u)\big)\right| \le C\mathrm{d}_{\mathrm{L}}(\mu_\Sigma,\mu_\sigma) \tag{3.18}$$

and

$$\left|\int_0^1 \big(m_{\Sigma,|w|}(\mathrm{i}\eta)-m_{\sigma,|w|}(\mathrm{i}\eta)\big)\,\mathrm{d}\eta\right| \le C\mathrm{d}_{\mathrm{L}}(\mu_\Sigma,\mu_\sigma), \tag{3.19}$$

uniformly for all $w \in \mathcal{R}_\sigma^\tau$, for N sufficiently large.

Inequality (3.18) follows from the continuity of the additive-free convolution. More precisely, from Theorem 4.13 of [12], we know that $\mathrm{d}_{\mathrm{L}}(\mu_{\Sigma,|w|},\mu_{\sigma,|w|}) \le \mathrm{d}_{\mathrm{L}}(\mu_\Sigma,\mu_\sigma)$. Since $\log|u-\mathrm{i}|$ is a smooth function and $\mu_{\Sigma,|w|}$, $\mu_{\sigma,|w|}$ are compactly supported, (3.18) follows.

To establish (3.19), we note that, for N sufficiently large,

$$\begin{aligned}\int_0^1 \big|m_{\Sigma,|w|}(\mathrm{i}\eta)-m_{\sigma,|w|}(\mathrm{i}\eta)\big|\,\mathrm{d}\eta &\le \max_{\eta\in(0,1]}\big|m_{\Sigma,|w|}(\mathrm{i}\eta)-m_{\sigma,|w|}(\mathrm{i}\eta)\big| \\ &\le C\mathrm{d}_{\mathrm{L}}(\mu_\Sigma,\mu_\sigma),\end{aligned}$$

for all w with $r_- + \tau \le |w| \le r_+ - \tau$, with a constant depending on τ. This follows directly from Theorem 2.7 of [2]. This shows (3.19).

So far we proved (1.20) for smooth functions f. Since ρ_σ is a Borel probability measure (see, e.g., Theorem 1.2), (1.20) extends to $f \in C^2(\mathbb{C})$ supported in $\mathcal{R}_\sigma^\tau$. This completes the proof of Corollary 1.12. □

4. Local law for block additive model. In this section we derive a local law for block additive random matrices in a slightly generalized setting; see Theorem 4.3 below. Theorem 2.5 is a direct consequence of this result.

First note that the matrix H^w defined in (2.14) can be rewritten as

$$H^w = \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix}\begin{pmatrix} 0 & \Sigma \\ \Sigma & 0 \end{pmatrix}\begin{pmatrix} U^* & 0 \\ 0 & V^* \end{pmatrix} + \begin{pmatrix} 0 & -w \\ -w^* & 0 \end{pmatrix}, \tag{4.1}$$

where 0 is the $N \times N$ matrix filled with zeros. In the following, we consider a slightly more general problem by looking at random matrices H defined by

$$H := \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix}\begin{pmatrix} 0 & \Sigma \\ \Sigma^* & 0 \end{pmatrix}\begin{pmatrix} U^* & 0 \\ 0 & V^* \end{pmatrix} + \begin{pmatrix} 0 & \Xi \\ \Xi^* & 0 \end{pmatrix}, \tag{4.2}$$

where

$$\Sigma := \mathrm{diag}(\sigma_1,\dots,\sigma_N), \qquad \Xi := \mathrm{diag}(\xi_1,\dots,\xi_N), \tag{4.3}$$

with $\sigma_i, \xi_i \in \mathbb{C}$, $i \in [\![1,N]\!]$. Here, Σ and Ξ are deterministic diagonal matrices, while U and V are independent Haar unitary or Haar orthogonal matrices of degree

N as before. Note that we allow in (4.3) for complex matrix elements in Σ and Ξ. In the sequel, we always assume that Σ and Ξ are bounded,

$$\|\Sigma\|, \|\Xi\| \le C, \tag{4.4}$$

for some constant C independent of N. Denote the empirical density of their singular values by

$$\mu_\Sigma := \frac{1}{N}\sum_{i=1}^{N} \delta_{|\sigma_i|}, \qquad \mu_\Xi := \frac{1}{N}\sum_{i=1}^{N} \delta_{|\xi_i|}. \tag{4.5}$$

Note that μ_Σ and μ_Ξ are probability measures on $[0, \infty)$. We assume that there are compactly supported probability measures μ_σ and μ_ξ such that

$$\sup_{N \ge N_0} \big(\mathrm{d_L}(\mu_\Sigma, \mu_\sigma) + \mathrm{d_L}(\mu_\Xi, \mu_\xi)\big) \le 2b, \tag{4.6}$$

for a sufficiently small constant $b > 0$ and sufficiently large N_0.

The following general regularity result is of interest.

LEMMA 4.1 (Theorem 4.1 in [8]). *Let μ_1 and μ_2 be Borel probability measures on $\mathbb{R}$, neither of them a point mass. Then the singular continuous part of $\mu_1 \boxplus \mu_2$ vanishes. A point $x \in \mathbb{R}$ is an atom of $\mu_1 \boxplus \mu_2$ if and only if there are $x_1, x_2 \in \mathbb{R}$ such that $x = x_1 + x_2$ and $\mu_1(\{x_1\}) + \mu_2(\{x_2\}) > 1$. Moreover, the absolutely continuous part of $\mu_1 \boxplus \mu_2$ is always nonzero, and its density is analytic wherever positive and finite.*

DEFINITION 4.2. For two Borel probability measures μ_1 on μ_2 on $\mathbb{R}$ satisfying the assumptions of Lemma 4.1, we set

$$\mathcal{B}_{\mu_1 \boxplus \mu_2} := \big\{x \in \mathbb{R} : 0 < f_{\mu_1 \boxplus \mu_2}(x) < \infty, \mu_1 \boxplus \mu_2(\{x\}) = 0\big\}, \tag{4.7}$$

where $f_{\mu_1 \boxplus \mu_2}$ denotes the density function of $\mu_1 \boxplus \mu_2$. We call $\mathcal{B}_\mu$ the bulk of μ.

Let $G \equiv G(z) := (H - z)^{-1}$ be the Green function of H at parameter $z \in \mathbb{C}^+$, and let

$$m_H(z) := \operatorname{tr} G(z) = \frac{1}{2N} \operatorname{Tr} G(z) \tag{4.8}$$

be the normalized trace of $G(z)$, which by the functional calculus agrees with the Stieltjes' transform of the empirical eigenvalue distribution of H.

Given an interval $\mathcal{I} \subset \mathbb{R}$ and $0 \le a \le b$, we introduce the domain

$$\mathcal{S}_{\mathcal{I}}(a, b) := \big\{z = E + \mathrm{i}\eta \in \mathbb{C}^+ : E \in \mathcal{I}, a < \eta \le b\big\}. \tag{4.9}$$

As before, we denote for a measure μ on $\mathbb{R}$ its symmetrization by μ^{sym}. The following is a key result of this paper.

THEOREM 4.3 (Strong law for H). *Suppose that (4.4) holds. Let μ_σ and μ_ξ be two compactly supported probability measures on $[0,\infty)$ such that neither $\mu_\sigma^{\mathrm{sym}}$ nor μ_ξ^{sym} is a single point mass and at least of one of them is supported at more than two points. Fix some $L > 0$ and let $\mathcal{I}$ be any compact interval of the bulk $\mathcal{B}_{\mu_\sigma^{\mathrm{sym}} \boxplus \mu_\xi^{\mathrm{sym}}}$. Then there exists a (small) constant $b_0 > 0$ and $N_0 \in \mathbb{N}$, depending only on μ_σ, μ_ξ, $\mathcal{I}$ and the constant C in (4.4), such that whenever*

$$\sup_{N \geq N_0} \big(\mathrm{d_L}(\mu_\Sigma, \mu_\sigma) + \mathrm{d_L}(\mu_\Xi, \mu_\xi)\big) \leq 2b, \tag{4.10}$$

for some $b \leq b_0$, then

$$\big|m_H(z) - m_{\mu_\Sigma^{\mathrm{sym}} \boxplus \mu_\Xi^{\mathrm{sym}}}(z)\big| \prec \frac{1}{N\eta(1+\eta)} \tag{4.11}$$

holds uniformly on $\mathcal{S}_{\mathcal{I}}(0, N^L)$, for N sufficiently large depending only on μ_σ, μ_ξ, $\mathcal{I}$, L and the constant C in (4.4). Moreover, there exists a constant $\eta_{\mathrm{M}} \geq 1$, independent of N, such that (4.11) holds uniformly on $\mathcal{S}_{\mathcal{I}}(\eta_M, N^L)$, for any compact interval $\mathcal{I} \subset \mathbb{R}$, for N sufficiently large depending only on μ_σ, μ_ξ, L and the constant C in (4.4).

Theorem 4.3 is proved in Sections 5–6 and Section 8 of the Supplementary Material [5]. In fact in [5], we prove Theorem 4.3 for spectral parameters $z \in \mathbb{C}^+$ with large imaginary parts, η. Here, large η means $\eta \geq \eta_{\mathrm{M}}$, for some $\eta_{\mathrm{M}} \geq 1$ independent of N to be chosen below. The proof for large η relies on the Gromov–Milman concentration inequality for the full Haar measure in conjunction with identities for expectations of the Green functions originating in the global $U(N)$-symmetry. These arguments are independent of the main line followed here and are hence postponed to Section 8 of the Supplementary Material [5]. The results for large η serve as initial estimates in a boostrap argument carried out in Sections 5–6 where we prove Theorem 4.3 in the complementary regime where $\eta < \eta_{\mathrm{M}}$.

PROOF OF THEOREM 2.5. Theorem 2.5 follows from Theorem 4.3 by choosing $\mathcal{I} = \{0\}$. The conditions of Theorem 4.3 require that the density of $\mu_\sigma^{\mathrm{sym}} \boxplus \mu_\xi^{\mathrm{sym}}$ is uniformly bounded from below on the compact interval $\mathcal{I}$. For $E = 0$, this condition was verified in Theorem 2.2. This yields (2.17) uniformly for $0 < \eta \leq N^L$, with $L > 1$ as in Theorem 4.3 for fixed w with $|w| \in [r_- + \tau, r_+ - \tau]$.

Next we show that (2.17) can be strengthened to a uniform bound in $w \in \mathcal{R}_\sigma^\tau := \{w \in \mathbb{C} : |w| \in [r_- + \tau, r_+ - \tau]\}$. We introduce the lattice

$$\widehat{\mathcal{R}}_\sigma^\tau(L_1) := \mathcal{R}_\sigma^\tau \cap N^{-L_1}\{\mathbb{Z} \times \mathrm{i}\mathbb{Z}\},$$

for some sufficiently large positive constant L_1 such that $L_1 \geq 2L$ (say). Using the definition of stochastic domination in Definition 1.6 and (2.17) for fixed w, we obtain

$$\max_{w \in \widehat{\mathcal{R}}_\sigma^\tau(L_1)} \big|m^w(z) - m_{\Sigma,|w|}(z)\big| \prec \frac{1}{N\eta},$$

uniformly in $0 < \eta \le N^L$. To extend this bound to all of $\mathcal{R}_\sigma^\tau$, it suffices to show Lipschitz continuity of these quantities in w. We need that, for any $w_1, w_2 \in \mathcal{R}_\sigma^\tau$ with $|w_1 - w_2| \le N^{-L_1}$ for sufficiently large L_1, one has

$$|m^{w_1}(z) - m^{w_2}(z)| \le \frac{1}{N\eta}, \qquad |m_{\Sigma,|w_1|}(z) - m_{\Sigma,|w_2|}(z)| \le \frac{1}{N\eta}, \tag{4.12}$$

uniformly in $0 < \eta \le N^L$. To show the first deterministic bound in (4.12), we use the bound

$$\begin{aligned} |m^{w_1}(z) - m^{w_2}(z)| &\le \frac{|w_1 - w_2|}{2N} \operatorname{Tr} |H^{w_1} - z|^{-1} |H^{w_2} - z|^{-1} \\ &\le \frac{|w_1 - w_2|}{2\eta^2} \le \frac{1}{2\eta} N^{-L_1+L} \le \frac{1}{N\eta}, \end{aligned}$$

where $|A| := \sqrt{A^* A}$, for any square matrix A.

To show the second bound in (4.12), we use the stability of the Stieltjes' transform of free additive convolution. Here it suffices to use the following bound (cf. (2.20) in [2] for instance):

$$|m_{\Sigma,|w_1|}(z) - m_{\Sigma,|w_2|}(z)| \le \frac{C}{\eta}\left(1 + \frac{1}{\eta}\right) \mathrm{d_L}(\delta^{\mathrm{sym}}_{|w_1|}, \delta^{\mathrm{sym}}_{|w_2|}) \le \frac{C}{\eta}\left(1 + \frac{1}{\eta}\right)|w_1 - w_2|,$$

for all $z = E + \mathrm{i}\eta \in \mathbb{C}^+$, where C is a constant uniform in z. Using the assumptions $|w_1 - w_2| \le N^{-L_1}$ and $L_1 \ge 2L$, we get (4.12), which in turn establishes the desired uniformity of (2.17) in $w \in \mathcal{R}_\sigma^\tau$.

To complete the proof of (2.17), it remains to deal with the large η regime, i.e., when $\eta \ge N^L$. For that we use the elementary (deterministic) estimates

$$m_H(\mathrm{i}\eta) = -\frac{1}{\mathrm{i}\eta} + O\left(\frac{1}{|\eta|^3}\right), \qquad m_{\Sigma,|w|}(\mathrm{i}\eta) = -\frac{1}{\mathrm{i}\eta} + O\left(\frac{1}{|\eta|^3}\right), \tag{4.13}$$

as $\eta \nearrow \infty$, where we used a resolvent expansion of G together with $\operatorname{tr} H = 0$ and $\|H\| \le S_+$ [see (1.3)], and the large η expansion of the Stieltjes' transform together with the fact that $\mu_{\Sigma,|w|}$ is symmetric and compactly supported. Thus for $\eta \ge N^L$, (2.17) follows from (4.13). Uniformity in $w \in \mathcal{R}_\sigma^\tau$ is immediate. □

4.1. *Approximate subordination for block additive models.* In this subsection, we establish the matrix subordination for the Green function of H. To simplify notation, we introduce the block matrices

$$A := \begin{pmatrix} 0 & \Xi \\ \Xi^* & 0 \end{pmatrix}, \qquad B := \begin{pmatrix} 0 & \Sigma \\ \Sigma^* & 0 \end{pmatrix}, \qquad \mathcal{U} := \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix}. \tag{4.14}$$

Then we write (4.2) as

$$H = A + \widetilde{B}, \qquad \widetilde{B} := \mathcal{U} B \mathcal{U}^*. \tag{4.15}$$

As before, we let $G(z) := (H - z)^{-1}$ be the Green function of H at spectral parameter $z \in \mathbb{C}^+$. A simple consequence of the definition of G are the identities

$$\widetilde{B}G(z) = I_{2N} - (A - z)G(z), \qquad G(z)\widetilde{B} = I_{2N} - G(z)(A - z). \tag{4.16}$$

Inspired by [32] (see also [3, 11, 29]), we introduce the approximate subordination functions

$$\omega_A^c(z) := z - \frac{\operatorname{tr} AG}{\operatorname{tr} G}, \qquad \omega_B^c(z) := z - \frac{\operatorname{tr} \widetilde{B}G}{\operatorname{tr} G}. \tag{4.17}$$

By these definitions and (4.16), we have

$$\omega_A^c(z) + \omega_B^c(z) - z = -\frac{1}{m_H(z)}. \tag{4.18}$$

Recall the measures μ_Σ and μ_Ξ of (4.5) as well as μ_σ and μ_ξ of (4.6). For their symmetrizations we introduce, hinting at (4.14), the shorthand

$$\mu_A \equiv \mu_\Xi^{\mathrm{sym}}, \qquad \mu_B \equiv \mu_\Sigma^{\mathrm{sym}}, \qquad \mu_\alpha \equiv \mu_\xi^{\mathrm{sym}}, \qquad \mu_\beta \equiv \mu_\sigma^{\mathrm{sym}}. \tag{4.19}$$

Note that μ_A and μ_B are the empirical spectral distributions of A and B. We denote by $\omega_A(z), \omega_B(z), \omega_\alpha(z), \omega_\beta(z)$ the subordination functions defined via (2.5) with the choices $(\mu_1, \mu_2) = (\mu_A, \mu_B)$ and (μ_α, μ_β), respectively.

The next result shows that the approximate subordination functions ω_A^c and ω_B^c are indeed good approximations to the subordination functions ω_A and ω_B. Moreover, it establishes the subordination for the diagonal Green function entries.

THEOREM 4.4. *Under the conditions and with the notation of Theorem 4.3, the estimates*

$$\left|\omega_A^c(z) - \omega_A(z)\right| \prec \frac{1}{N\eta}, \qquad \left|\omega_B^c(z) - \omega_B(z)\right| \prec \frac{1}{N\eta}, \tag{4.20}$$

hold uniformly on $\mathcal{S}_{\mathcal{I}}(0, \eta_{\mathrm{M}})$, for N sufficiently large depending only on μ_α, μ_β, $\mathcal{I}$, L and the constant C in (4.4). Moreover, we have

$$\begin{aligned}
\left|G_{ii}(z) - \frac{\omega_B(z)}{|\xi_i|^2 - (\omega_B(z))^2}\right| &\prec \frac{1}{\sqrt{N\eta}},\\
\left|G_{\hat{i}\hat{i}(z)} - \frac{\omega_B(z)}{|\xi_i|^2 - (\omega_B(z))^2}\right| &\prec \frac{1}{\sqrt{N\eta}},\\
\left|G_{i\hat{i}}(z) - \frac{\xi_i}{|\xi_i|^2 - (\omega_B(z))^2}\right| &\prec \frac{1}{\sqrt{N\eta}},\\
\left|G_{\hat{i}i}(z) - \frac{\bar{\xi}_i}{|\xi_i|^2 - (\omega_B(z))^2}\right| &\prec \frac{1}{\sqrt{N\eta}},
\end{aligned} \tag{4.21}$$

uniformly in $i \in [\![1, N]\!]$ and in $z \in \mathcal{S}_{\mathcal{I}}(0, \eta_{\mathrm{M}})$, where $\hat{i} := i + N$, for N sufficiently large depending only on μ_α, μ_β, $\mathcal{I}$, L and the constant C in (4.4).

REMARK 4.5. Some crucial properties of the subordination functions ω_A and ω_B are collected in Lemma A.2 in the Supplementary Material [5]. Here, we mention that under the assumptions of Theorem 4.4, for N sufficiently large, the imaginary parts of the subordination functions, $\operatorname{Im}\omega_A(z)$ and $\operatorname{Im}\omega_B(z)$ are both bounded from below on $z \in \mathcal{S}_{\mathcal{I}}(0, \eta_{\mathrm{M}})$. This follows from Lemma A.2 and the assumption that $\mathcal{I}$ is a compact interval in the bulk of $\mu_\alpha \boxplus \mu_\beta$. It then follows from (4.21) that $|G_{ii}(z)| \prec 1$ and $|G_{\hat{i}\hat{i}}(z)| \prec 1$ uniformly on $\mathcal{S}_{\mathcal{I}}(N^{-1+\gamma}, \eta_{\mathrm{M}})$, for any $\gamma > 0$, and all $i \in [\![1, N]\!]$. A direct consequence of this result is that the eigenvectors associated with eigenvalues in the bulk are fully delocalized. More precisely, letting $(\boldsymbol{u}_k)$ denote the ℓ^2-normalized eigenvectors associated with the eigenvalues (λ_k), $k \in [\![1, 2N]\!]$, we have

$$\max_{k:\lambda_k \in \mathcal{I}} \|\boldsymbol{u}_k\|_\infty \prec \frac{1}{\sqrt{N}}, \tag{4.22}$$

for any compact interval $\mathcal{I}$ in the bulk of $\mu_\alpha \boxplus \mu_\beta$. For a proof of (4.22) from Theorem 4.4, we refer to the proof of Theorem 2.6 in [3].

4.2. *Outline of the strategy of proof.* The proof of the local law of Theorem 4.3 is carried out in three steps. In Step 1, we consider the large η regime, that is, we establish (4.11) on $\mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{M}}, N^L)$, for some sufficiently large, but N-independent, η_{M}. In Step 2, we establish a weak local law for m^w in the small η regime, that is, we establish (4.11) with a weaker error bound on $\mathcal{S}_{\mathcal{I}}(N^{-1+\gamma}, \eta_{\mathrm{M}})$, for some small $\gamma > 0$; see Theorem 5.1 below for the statement of the weak law. The extension to $\mathcal{S}_{\mathcal{I}}(0, \eta_{\mathrm{M}})$ will follow directly from monotonicity of the Green function. This second step is based on a bootstrapping argument to reduce the spectral parameter $\operatorname{Im} z$. Step 1 will provide the initial estimate to get the bootstrapping started. In Step 3, we use a fluctuation averaging argument together with the weak local law established in the second step to get (4.11) in its strong form.

Step 1 is carried out in Section 8. It builds on the celebrated Gromov–Milman concentration inequality whose application to random matrix theory is fairly standard [1]. For additive models of the form $X + UYU^*$, with deterministic $X, Y \in M_N(\mathbb{C})$ and U Haar distributed on $U(N)$ or on $O(N)$ it was used in [2, 29, 32], and for the model block-additive model considered in this section in [11].

Step 2 is carried out in Section 5, where we prove Theorem 5.1. This proof has three major ingredients. First we use a partial randomness decomposition of the Haar measure [see (5.2)] that enables us to take partial expectations of functions of the diagonal Green function entries G_{ii}, $G_{\hat{i}\hat{i}}$. Exploiting concentration only for this partial randomness surpasses the more general but less flexible Gromov–Milman technique used in Step 1. Second, to compute the partial expectations of G_{ii}, we establish a system of self-consistent equations involving only two auxiliary quantities (S_{ij}) and (T_{ij}); see (5.16). In our previous work [3], we used a similar approach to derive the local law for $X + UYU^*$. For the model considered

in this paper, we face with a new phenomenon causing several substantial difficulties. The main point is that for block additive models, we have less randomness originating in the Haar measure on $U(N) \times U(N)$ than for the additive models with Haar measure on $U(2N)$. As a consequence, we have to control more quantities in the two blocks separately. Even more importantly, the coupling between the two blocks is provided solely by the diagonal matrix Σ without any randomness; see (4.2). Our proof shows that the randomness in the diagonal blocks and the deterministic off-diagonal blocks effectively make up for the lacking off-diagonal randomness.

To derive the aforementioned system of equations for (S_{ij}), (T_{ij}) and (G_{ii}), we use the partial decomposition of Haar measure in combination with recursive moment estimates; see, for example, Lemma 5.3 for such a statement. Recursive moment estimates were used first in [30] to derive local laws for sparse Wigner matrices. They allow us to pass on cumbersome partial concentration estimates used in Section 5 of [3], and provide a conceptually clear approach to the weak local law for both models. Third, to connect the diagonal Green function entries with the subordination functions from Theorem 2.1, we rely on the optimal stability result for the subordination equations obtained in [2].

Step 3 is carried out in Section 6. In this section we exploit the so-called fluctuation averaging mechanism to improve the estimates of Step 2. While the fluctuation averaging mechanism is, thanks to the independence of the matrix entries, well understood for Wigner type matrices (see e.g., [20, 22]), dependencies among the entries of the Haar matrices mask this mechanism and its current understanding for matrix ensembles involving Haar matrices is still rather poor. We gave a first result in [4] for additive models. In the present paper, we approach the fluctuation mechanics for block-additive models by first deriving a set of so-called "Ward identities" which will enable us to complete the proof of Theorem 4.3. Ward identities are relations among tracial quantities involving the Green function and the matrices A and B. In expectation, these relations can be derived using the invariance of Haar measure (see, e.g., (8.11) for a first example), yet we will require optimal estimates that hold with high probability; see, for example, (5.21) and (6.3). These estimates are obtained using recursive moment estimates for carefully chosen quantities; see (5.19). Since we have less randomness coming from $U(N) \times U(N)$ in the setup of block-additive models, more quantities need to be simultaneously controlled than in the additive models, resulting in a more sophisticated analysis.

4.3. *Notation*. We introduce some more notation used in the proof of Theorem 4.3.

Notation for matrices: In our analysis, we also use the matrices

$$\mathcal{H} = B + \mathcal{U}^* A \mathcal{U} =: B + \widetilde{A}, \qquad \mathcal{G}(z) = (\mathcal{H} - z)^{-1}, \tag{4.23}$$

which are the analogues of H in (4.15) and of its Green function $G(z)$, obtained by switching the roles of A and B, and also the roles of $\mathcal{U}$ and $\mathcal{U}^*$. Note that by cyclicity $\operatorname{Tr} G(z) = \operatorname{Tr} \mathcal{G}(z)$.

Vector space notation: For any index $i \in [\![1, N]\!]$, we let $\hat{i} \equiv i + N$. We make the convention hereafter that the index i always runs from 1 to N, unless said otherwise. Thus the index $\hat{i}$ runs from $N+1$ to $2N$. We denote by $\sum_i^{(k)}$ the sum over $i \in [\![1, N]\!] \setminus \{k\}$. We denote by $\{\boldsymbol{e}_i\}$ the canonical basis of $\mathbb{C}^N$ while we denote by $\{\hat{\boldsymbol{e}}_i\}$ the canonical basis of $\mathbb{C}^{2N}$. We let $\mathbf{0}$ denote the zero vector in either space. We use bold font for vectors and denote the components as $\boldsymbol{v} = (v_i)$.

The identity matrix in $M_N(\mathbb{C})$, respectively, $M_{2N}(\mathbb{C})$, is denoted by

$$I \equiv I_N, \qquad \hat{I} \equiv I_{2N}, \tag{4.24}$$

and we let

$$\hat{I}_1 := I \oplus 0, \qquad \hat{I}_2 := 0 \oplus I \tag{4.25}$$

denote the block identities in $M_N(\mathbb{C}) \oplus M_N(\mathbb{C})$, where 0 represents the $N \times N$ zero matrix.

For any matrix $D \in M_n(\mathbb{C})$, $n \geq 1$, we let

$$\operatorname{tr} D := \frac{1}{n} \operatorname{Tr} D$$

denote the normalized trace of D. For $D \in M_{2N}(\mathbb{C})$, we introduce the normalized partial traces

$$\tau_1(D) := \frac{1}{N} \sum_{i=1}^{N} D_{ii}, \qquad \tau_2(D) := \frac{1}{N} \sum_{i=1}^{N} D_{\hat{i}\hat{i}}. \tag{4.26}$$

Using the block structure of H, it is easy to check that the Green function $G(z)$ satisfies

$$\tau_1\big(G(z)\big) = \tau_2\big(G(z)\big), \qquad z \in \mathbb{C}^+. \tag{4.27}$$

Φ-system: For our purposes, it is convenient to recast (2.5) in a compact form: For generic probability measures μ_1, μ_2 on $\mathbb{R}$, let the function $\Phi_{\mu_1,\mu_2} : (\mathbb{C}^+)^3 \to \mathbb{C}^2$ be given by

$$\Phi_{\mu_1,\mu_2}(\omega_1, \omega_2, z) := \begin{pmatrix} F_{\mu_1}(\omega_2) - \omega_1 - \omega_2 + z \\ F_{\mu_2}(\omega_1) - \omega_1 - \omega_2 + z \end{pmatrix}. \tag{4.28}$$

Considering μ_1, μ_2 as fixed, the equation

$$\Phi_{\mu_1,\mu_2}(\omega_1, \omega_2, z) = 0, \tag{4.29}$$

is equivalent to (2.5) and, by Theorem 2.1, there are unique analytic functions $\omega_1, \omega_2 : \mathbb{C}^+ \to \mathbb{C}^+$, $z \mapsto \omega_1(z), \omega_2(z)$ satisfying (2.4) that solve (4.29) in terms of z.

Control parameters: For $z \in \mathbb{C}^+$, we will use the following deterministic control parameter:

$$\Psi \equiv \Psi(z) := \frac{1}{\sqrt{N\eta(1+\eta)}}, \qquad \eta = \operatorname{Im} z. \tag{4.30}$$

We further introduce, for $z \in \mathbb{C}^+$ and $i \in [\![1, N]\!]$, the random control parameters

$$\begin{aligned}
\Lambda_{\mathrm{d};ii}(z) &:= \left| G_{ii} - \frac{\omega_B(z)}{|\xi_i|^2 - (\omega_B(z))^2} \right|, \\
\Lambda_{\mathrm{d};\hat{i}\hat{i}}(z) &:= \left| G_{\hat{i}\hat{i}} - \frac{\omega_B(z)}{|\xi_i|^2 - (\omega_B(z))^2} \right|, \\
\Lambda_{\mathrm{d};i\hat{i}}(z) &:= \left| G_{i\hat{i}} - \frac{\xi_i}{|\xi_i|^2 - (\omega_B(z))^2} \right|, \\
\Lambda_{\mathrm{d};\hat{i}i}(z) &:= \left| G_{\hat{i}i} - \frac{\bar{\xi}_i}{|\xi_i|^2 - (\omega_B(z))^2} \right|, \\
\Lambda_{\mathrm{d}}(z) &:= \max_{i \in [\![1,N]\!]} \max_{k,l=i \text{ or } \hat{i}} \Lambda_{\mathrm{d};kl}(z).
\end{aligned} \tag{4.31}$$

We also define $\Lambda_{\mathrm{d}}^c(z)$ analogously by replacing ω_B by ω_B^c (cf. (4.17)) in the definition of $\Lambda_{\mathrm{d}}(z)$. We will often omit the variable z from the above notation when there is no confusion.

For notational simplicity, we do not follow the threshold N for which the estimates apply. Following the dependence of this threshold on the other parameters along the proofs, one may easily verify the dependences stated in Theorem 4.3 and Theorem 4.4.

5. Green function subordination for small η. Let $\eta_{\mathrm{M}} > 0$ be some sufficiently large constant, and for any given (small) $\gamma > 0$, we set

$$\eta_{\mathrm{m}} \equiv \eta_{\mathrm{m}}(\gamma) := N^{-1+\gamma}. \tag{5.1}$$

In this section we prove a Green function subordination property in the regime $\eta_{\mathrm{m}} \le \eta \le \eta_{\mathrm{M}}$. The formal statement is given in Theorem 5.1 below. For definiteness, we work with the unitary setup in this section. The necessary modifications for the orthogonal case are stated in Appendix C of the Supplementary Material [5]. We start with the partial randomness decomposition of the Haar measure on $U(N) \times U(N)$ announced in Section 4.2.

5.1. *Partial randomness decomposition of the Haar measure.* Let $\boldsymbol{u}_i = (u_{i1}, \dots, u_{iN})'$ and $\boldsymbol{v}_i = (v_{i1}, \dots, v_{iN})'$ be the ith columns of U and V, respectively. Let θ_i^u and θ_i^v be the arguments of u_{ii} and v_{ii}, respectively, and let $\phi_i^a = \mathrm{e}^{\mathrm{i}\theta_i^a}$ for $a = u, v$. Our approach relies on the partial randomness decomposition of the Haar measure from [18, 31]:

$$U = -\phi_i^u R_i^u U^{\langle i \rangle}, \qquad V = -\phi_i^v R_i^v V^{\langle i \rangle}. \tag{5.2}$$

Here, $U^{\langle i\rangle}$ and $V^{\langle i\rangle}$ are unitary matrices with (i,i)th entry equal 1, and their (i,i)-minors are independent, Haar distributed on $\mathcal{U}(N-1)$. In particular, $U^{\langle i\rangle}\boldsymbol{e}_i = V^{\langle i\rangle}\boldsymbol{e}_i = \boldsymbol{e}_i$ and $\boldsymbol{e}_i^* U^{\langle i\rangle} = \boldsymbol{e}_i^* V^{\langle i\rangle} = \boldsymbol{e}_i^*$, where $\boldsymbol{e}_i$ is the ith coordinate vector. In addition, $U^{\langle i\rangle}$ is independent of $\boldsymbol{u}_i$, and $V^{\langle i\rangle}$ is independent of $\boldsymbol{v}_i$. Here R_i^u and R_i^v are reflections, defined as

$$R_i^a := I - \boldsymbol{r}_i^a (\boldsymbol{r}_i^a)^*, \qquad a = u, v, \tag{5.3}$$

where

$$\boldsymbol{r}_i^u := \sqrt{2}\frac{\boldsymbol{e}_i + \bar{\phi}_i^u \boldsymbol{u}_i}{\|\boldsymbol{e}_i + \bar{\phi}_i^u \boldsymbol{u}_i\|_2}, \qquad \boldsymbol{r}_i^v := \sqrt{2}\frac{\boldsymbol{e}_i + \bar{\phi}_i^v \boldsymbol{v}_i}{\|\boldsymbol{e}_i + \bar{\phi}_i^v \boldsymbol{v}_i\|_2}. \tag{5.4}$$

Note that R_i^u is independent of $U^{\langle i\rangle}$ and R_i^v is independent of $V^{\langle i\rangle}$.

Set the $(2N)\times(2N)$ matrices

$$\Phi_i := (\phi_i^u I) \oplus (\phi_i^v I), \qquad \mathcal{R}_i := R_i^u \oplus R_i^v, \qquad \mathcal{U}_i := U^{\langle i\rangle} \oplus V^{\langle i\rangle}. \tag{5.5}$$

With the above notation and the decompositions in (5.2), we have

$$\mathcal{U} = -\mathcal{R}_i \mathcal{U}_i \Phi_i. \tag{5.6}$$

Hence, for each $i \in [\![1, N]\!]$, we can write

$$H = A + \widetilde{B} = A + \mathcal{R}_i \mathcal{U}_i \Phi_i B \Phi_i^* \mathcal{U}_i^* \mathcal{R}_i := A + \mathcal{R}_i \widetilde{B}^{\langle i\rangle} \mathcal{R}_i, \tag{5.7}$$

where we introduced the notation

$$\widetilde{B}^{\langle i\rangle} := \mathcal{U}_i \Phi_i B \Phi_i^* \mathcal{U}_i^*. \tag{5.8}$$

We further define the matrices

$$H^{\langle i\rangle} := A + \widetilde{B}^{\langle i\rangle}, \qquad G^{\langle i\rangle} := \big(H^{\langle i\rangle} - z\big)^{-1}. \tag{5.9}$$

Since $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are independent, uniformly distributed complex unit vectors, there exist independent normal vectors, $\widetilde{\boldsymbol{g}}_i^u, \widetilde{\boldsymbol{g}}_i^v \sim \mathcal{N}_{\mathbb{C}}(0, \frac{1}{N} I_N)$ such that

$$\boldsymbol{u}_i = \frac{\widetilde{\boldsymbol{g}}_i^u}{\|\widetilde{\boldsymbol{g}}_i^u\|_2}, \qquad \boldsymbol{v}_i = \frac{\widetilde{\boldsymbol{g}}_i^v}{\|\widetilde{\boldsymbol{g}}_i^v\|_2}.$$

We further define

$$\boldsymbol{g}_i^u := \bar{\phi}_i^u \widetilde{\boldsymbol{g}}_i^u, \qquad \boldsymbol{h}_i^u := \frac{\boldsymbol{g}_i^u}{\|\boldsymbol{g}_i^u\|_2} = \bar{\phi}_i^u \boldsymbol{u}_i, \qquad \ell_i^u := \frac{\sqrt{2}}{\|\boldsymbol{e}_i + \boldsymbol{h}_i^u\|_2}, \tag{5.10}$$

and define $\boldsymbol{g}_i^v$, $\boldsymbol{h}_i^v$ and ℓ_i^v analogously by replacing $\boldsymbol{u}_i$ by $\boldsymbol{v}_i$. Note that for $a = u$ or v, g_{ik}^a's for $k \neq i$ are $N_{\mathbb{C}}(0, \frac{1}{N})$ variables and g_{ii}^a is χ-distributed with $\mathbb{E}[(g_{ii}^a)^2] = \frac{1}{N}$. In addition, the components of $\boldsymbol{g}_i^a$ are independent, and they are all independent of ϕ_i^a. Hence, $\boldsymbol{g}_i^a$ and $\boldsymbol{h}_i^a$ are independent of $\widetilde{B}^{\langle i\rangle}$ (cf. (5.8)), for $a = u, v$. With this notation we can write

$$\boldsymbol{r}_i^a = \ell_i^a (\boldsymbol{e}_i + \boldsymbol{h}_i^a), \qquad a = u, v, \tag{5.11}$$

where $\boldsymbol{r}_i^a$ is defined in (5.4). Using Lemma A.1, it is elementary to check that, for $a = u, v$,

$$\begin{aligned} \|\boldsymbol{g}_i^a\|_2 &= 1 + \frac{1}{2}(\|\boldsymbol{g}_i^a\|_2^2 - 1) + O_{\prec}\Big(\frac{1}{N}\Big), \\ (\ell_i^a)^2 &= \frac{1}{1 + \boldsymbol{e}_i^* \boldsymbol{h}_i^a} = 1 - g_{ii}^a + O_{\prec}\Big(\frac{1}{N}\Big), \end{aligned} \tag{5.12}$$

where in the first estimate we used the fact $|\|\boldsymbol{g}_i^a\|_2^2 - 1| \prec \frac{1}{\sqrt{N}}$. In addition, by definition, R_i^a is a reflection sending $\boldsymbol{e}_i$ to $-\boldsymbol{h}_i^a$, that is,

$$R_i^a \boldsymbol{e}_i = -\boldsymbol{h}_i^a, \qquad R_i^a \boldsymbol{h}_i^a = -\boldsymbol{e}_i, \qquad a = u, v. \tag{5.13}$$

We also denote by $\mathring{\boldsymbol{g}}_i^a$ the vector obtained from $\boldsymbol{g}_i^a$ by replacing g_{ii}^a by 0, that is,

$$\mathring{\boldsymbol{g}}_i^a := \boldsymbol{g}_i^a - g_{ii}^a \boldsymbol{e}_i, \qquad a = u, v.$$

Correspondingly, we set

$$\mathring{\boldsymbol{h}}_i^a := \frac{\mathring{\boldsymbol{g}}_i^a}{\|\boldsymbol{g}_i^a\|_2}, \qquad a = u, v. \tag{5.14}$$

Recall the notation $\mathbf{0}$ for the $N \times 1$ null vector. Finally, for brevity, we set

$$\boldsymbol{k}_i^u := \begin{pmatrix} \boldsymbol{h}_i^u \\ \mathbf{0} \end{pmatrix}, \qquad \boldsymbol{k}_i^v := \begin{pmatrix} \mathbf{0} \\ \boldsymbol{h}_i^v \end{pmatrix}, \qquad \mathring{\boldsymbol{k}}_i^u := \begin{pmatrix} \mathring{\boldsymbol{h}}_i^u \\ \mathbf{0} \end{pmatrix}, \qquad \mathring{\boldsymbol{k}}_i^v := \begin{pmatrix} \mathbf{0} \\ \mathring{\boldsymbol{h}}_i^v \end{pmatrix}. \tag{5.15}$$

We move on to the formal statement of the Green function subordination.

5.2. *Green function subordination.* Recall the notation $\{\hat{\boldsymbol{e}}_i\}$ for the standard basis of $\mathbb{C}^{2N}$, and also the notation $\hat{i} \equiv i + N$ for any $i \in [\![1, N]\!]$. We introduce the following quantities for $j = i, \hat{i}$, $i \in [\![1, N]\!]$,

$$\begin{aligned} S_{ij} &:= (\boldsymbol{k}_i^u)^* \widetilde{B}^{\langle i \rangle} G \hat{\boldsymbol{e}}_j, \qquad T_{ij} := (\boldsymbol{k}_i^u)^* G \hat{\boldsymbol{e}}_j, \\ S_{\hat{i}j} &:= (\boldsymbol{k}_i^v)^* \widetilde{B}^{\langle i \rangle} G \hat{\boldsymbol{e}}_j, \qquad T_{\hat{i}j} := (\boldsymbol{k}_i^v)^* G \hat{\boldsymbol{e}}_j \end{aligned} \tag{5.16}$$

and

$$\begin{aligned} \mathring{S}_{ii} &:= (\mathring{\boldsymbol{k}}_i^u)^* \widetilde{B}^{\langle i \rangle} G \hat{\boldsymbol{e}}_i = S_{ii} - \tilde{\sigma}_i h_{ii}^u G_{\hat{i}i}, \\ \mathring{T}_{ii} &:= (\mathring{\boldsymbol{k}}_i^u)^* G \hat{\boldsymbol{e}}_i = T_{ii} - h_{ii}^u G_{ii}, \end{aligned} \tag{5.17}$$

where $\tilde{\sigma}_i = \phi_i^u \bar{\phi}_i^v \sigma_i$, and σ_i is the ith diagonal entry of Σ; cf. (4.3). Here, in (5.17) we used

$$\hat{\boldsymbol{e}}_i^* \widetilde{B}^{\langle i \rangle} = \tilde{\sigma}_i \hat{\boldsymbol{e}}_{\hat{i}}^*, \qquad \widetilde{B}^{\langle i \rangle} \hat{\boldsymbol{e}}_{\hat{i}} = \tilde{\sigma}_i \hat{\boldsymbol{e}}_i, \qquad i \in [\![1, N]\!], \tag{5.18}$$

which is checked from the definitions of $\widetilde{B}^{\langle i \rangle}$ in (5.8), $\mathcal{U}_i$ and Φ_i in (5.5), and also B in (4.14).

Recall from (4.26) the notations for normalized partial traces τ_1 and τ_2 on $M_{2N}(\mathbb{C})$. Moreover, recall from (4.31) the definition of the control parameters $\Lambda_{\mathrm{d};ii}(z)$, $\Lambda_{\mathrm{d};\hat{i}\hat{i}}(z)$, $\Lambda_{\mathrm{d};i\hat{i}}(z)$, $\Lambda_{\mathrm{d};\hat{i}i}(z)$ and $\Lambda_{\mathrm{d}}(z)$. We further introduce $\Lambda_{\mathrm{d}}^c(z)$ analogously by replacing ω_B by ω_B^c (cf. (4.17)) in the definition of $\Lambda_{\mathrm{d}}(z)$. We will often omit the variable z from this notation.

In this section we will show that $\Lambda_{\mathrm{d}}(z)$, $\Lambda_{\mathrm{d}}^c(z)$ and Λ_T are of order Ψ with high probability; that is, matrix elements of the Green function can be expressed in terms of the subordination functions, up to a small random fluctuations of order Ψ. We will refer to these results as *Green function subordination*. The main tool is a high moment calculation and Gaussian integration by parts. However, we cannot directly estimate the high moments of T_{kl} and the formulas $|G_{ij} - [\ldots]|$ defining $\Lambda_{\mathrm{d};ij}(z)$. Instead, we introduce the following auxiliary quantities. For each $i \in [\![1, N]\!]$ and $j = i$ or $\hat{i}$, let

$$\begin{aligned}
\mathcal{P}_{ij} &\equiv \mathcal{P}_{ij}(z) := (\widetilde{B}G)_{ij}\tau_1(G) - G_{ij}\tau_1(\widetilde{B}G) + (G_{ij} + T_{ij})\Upsilon_1,\\
\mathcal{P}_{\hat{i}j} &\equiv \mathcal{P}_{\hat{i}j}(z) := (\widetilde{B}G)_{\hat{i}j}\tau_2(G) - G_{\hat{i}j}\tau_2(\widetilde{B}G) + (G_{\hat{i}j} + T_{\hat{i}j})\Upsilon_2,\\
\mathcal{K}_{ij} &\equiv \mathcal{K}_{ij}(z) := T_{ij} + \tau_1(G)\big(\tilde{\sigma}_i T_{\hat{i}j} + (\widetilde{B}G)_{ij}\big) - \tau_1(G\widetilde{B})(G_{ij} + T_{ij}),\\
\mathcal{K}_{\hat{i}j} &\equiv \mathcal{K}_{\hat{i}j}(z) := T_{\hat{i}j} + \tau_2(G)\big(\tilde{\sigma}_i^* T_{ij} + (\widetilde{B}G)_{\hat{i}j}\big) - \tau_2(G\widetilde{B})(G_{\hat{i}j} + T_{\hat{i}j}),
\end{aligned} \tag{5.19}$$

where, with $a = 1, 2$,

$$\Upsilon_a \equiv \Upsilon_a(z) := \tau_a(\widetilde{B}G) + \tau_a(G)\tau_a(\widetilde{B}G\widetilde{B}) - \tau_a(G\widetilde{B})\tau_a(\widetilde{B}G). \tag{5.20}$$

Using the invariance of the Haar measure, the following Ward identities

$$\mathbb{E}\Upsilon_a = 0, \qquad a = 1, 2, \tag{5.21}$$

can be checked. However, we will also need to know that Υ_a are small with high probability and not only in expectation in the following; see, for example, (5.29) in Theorem 5.2 below.

We will compute their high moments of these auxiliary quantities $\mathcal{P}$ and $\mathcal{K}$ and from them we will conclude the estimates on the Λ's. The careful choice of these auxiliary quantities $\mathcal{P}$ and $\mathcal{K}$ is essential for the proof. They have a built-in cancellation mechanism that makes the high moment calculation tractable; see (5.53)–(5.55) later.

Moreover, we recall the following matrices introduced in (4.23)

$$\mathcal{H} = B + \mathcal{U}^* A \mathcal{U} =: B + \widetilde{A}, \qquad \mathcal{G}(z) = (\mathcal{H} - z)^{-1}, \qquad z \in \mathbb{C}^+,$$

which are the analogue of H in (4.15) and its Green function $G(z)$, obtained via swapping the roles of A and B, and also the roles of $\mathcal{U}$ and $\mathcal{U}^*$. Note that the structure of $\mathcal{H}$ is exactly the same as H, so we can define the $\mathcal{H}$-counterparts of all quantities we have introduced so far for H. We will not repeat the heavy notation of the partial randomness decomposition for $\mathcal{H}$ as well, since we will not need all

these details. We will only need to know that, accordingly, we can define $\mathcal{G}_{ij}$, $\mathcal{S}_{ij}$ and $\mathcal{T}_{ij}$ by applying the same switching in the definitions of G_{ij}, S_{ij} and T_{ij}.

Also note the following alternative definition of ω_A^c and ω_B^c in (4.17):

$$\omega_A^c(z) := z - \frac{\operatorname{tr} \widetilde{A}\mathcal{G}}{\operatorname{tr}\mathcal{G}}, \qquad \omega_B^c(z) := z - \frac{\operatorname{tr} B\mathcal{G}}{\operatorname{tr}\mathcal{G}}, \tag{5.22}$$

and the trivial fact $\operatorname{tr} G = \operatorname{tr}\mathcal{G}$.

In addition, replacing ξ_i, ω_B, G_{ij} by $\sigma_i, \omega_A, \mathcal{G}_{ij}$ respectively in (4.31), we define $\widetilde{\Lambda}_{\mathrm{d};ij}(z)$ and $\widetilde{\Lambda}_{\mathrm{d}}(z)$ as the analogues of $\Lambda_{\mathrm{d};ij}(z)$ and $\Lambda_{\mathrm{d}}(z)$. For example,

$$\widetilde{\Lambda}_{\mathrm{d};ii}(z) = \left| \mathcal{G}_{ii} - \frac{\omega_A(z)}{|\sigma_i|^2 - (\omega_A(z))^2} \right| \tag{5.23}$$

and

$$\widetilde{\Lambda}_{\mathrm{d}}(z) := \max_{i \in [\![1,N]\!]} \max_{k,l=i \text{ or } \hat{i}} \widetilde{\Lambda}_{\mathrm{d};kl}(z). \tag{5.24}$$

Similarly, we can also define $\widetilde{\Lambda}_{\mathrm{d}}^c(z)$ and $\widetilde{\Lambda}_T(z)$ as the analogue of $\Lambda_{\mathrm{d}}^c(z)$ and $\Lambda_T(z)$, respectively. The analysis of the operator $\mathcal{H}$ is very similar to that of H, but at some point it will be useful to work with them in tandem, so we will need to control both.

Our main aim in this section is to prove the following Green function subordination property. Recall the definition of the control parameter $\Psi(z)$ from (4.30).

THEOREM 5.1. *Suppose that the assumptions in Theorem 4.3 hold. Then*

$$\begin{aligned} &\Lambda_{\mathrm{d}}(z) \prec \Psi(z), \qquad &&\widetilde{\Lambda}_{\mathrm{d}}(z) \prec \Psi(z), \\ &\Lambda_T(z) \prec \Psi(z), \qquad &&\widetilde{\Lambda}_T(z) \prec \Psi(z) \end{aligned} \tag{5.25}$$

uniformly on $\mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$, *for any* (*large*) *constant* $\eta_{\mathrm{M}} > 0$ *and* (*small*) *constant* $\gamma > 0$, *in the definition of* η_{m} (*cf.* (5.1)). *Moreover, the estimates*

$$\begin{aligned} &\left|\omega_A^c(z) - \omega_A(z)\right| \prec \Psi(z), \qquad &&\left|\omega_B^c(z) - \omega_B(z)\right| \prec \Psi(z), \\ &\left|m_H(z) - m_{\mu_A \boxplus \mu_B}(z)\right| \prec \Psi(z) \end{aligned} \tag{5.26}$$

also hold uniformly on $\mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$.

The estimates on the tracial quantities and the subordination functions in (5.26) are weaker than the final result in Theorem 4.3 and Theorem 4.4. Later in Section 6, we will improve them. The estimates in (5.25) are, however, (believed to be) optimal.

In what follows, we will mainly work with $\Lambda_{\mathrm{d}}(z)$. The discussion on $\widetilde{\Lambda}_{\mathrm{d}}(z)$ is the same. First we show the analogous estimate for Λ_{d}^c by assuming an a priori bound on Λ_{d} and Λ_T, for a fixed $z \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$. This is the content of Theorem 5.2 below. A continuity argument in Section 5.5 then allows us to conclude Theorem 5.1 from Theorem 5.2.

THEOREM 5.2. *Suppose that the assumptions in Theorem 4.3 hold. Let $\eta_{\mathrm{M}} > 0$ be a (large) constant and $\gamma > 0$ be a (small) constant in (5.1). Fix a $z \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$. Assume that*

$$\Lambda_{\mathrm{d}}(z) \prec N^{-\frac{\gamma}{4}}, \qquad \widetilde{\Lambda}_{\mathrm{d}}(z) \prec N^{-\frac{\gamma}{4}}, \qquad \Lambda_T(z) \prec 1, \qquad \widetilde{\Lambda}_T(z) \prec 1. \tag{5.27}$$

Then we have

$$\begin{aligned} &|\mathcal{P}_{ij}(z)| \prec \Psi(z), \qquad |\mathcal{P}_{\hat{i}j}(z)| \prec \Psi(z), \\ &|\mathcal{K}_{ij}(z)| \prec \Psi(z), \qquad |\mathcal{K}_{\hat{i}j}(z)| \prec \Psi(z), \end{aligned} \tag{5.28}$$

for all $i \in [\![1, N]\!]$ and $j = i$ or $\hat{i}$. In addition, under (5.27) we also have

$$|\Upsilon_1(z)| \prec \Psi(z), \qquad |\Upsilon_2(z)| \prec \Psi(z) \tag{5.29}$$

and

$$\Lambda_{\mathrm{d}}^{c}(z) \prec \Psi(z), \qquad \Lambda_T(z) \prec \Psi(z). \tag{5.30}$$

The same statements hold if we switch the roles of A and B, and also the roles of U and U^*, in all the conclusions from (5.28) to (5.30).

Note that, since $\eta_{\mathrm{m}} \le \eta \le \eta_{\mathrm{M}}$, we have $\Psi(z) \sim \frac{1}{\sqrt{N\eta}}$.

The proof of Theorem 5.2 proceeds in two steps. In the first step, we establish in Section 5.3 recursive moment estimates for the quantities $\mathcal{P}_{ii}$ and $\mathcal{K}_{ii}$. In the second step, carried out in Section 5.4, we use a local stability analysis to conclude Theorem 5.2 from the estimates established in Section 5.3.

5.3. *Recursive moment estimates for $\mathcal{P}_{ii}$ and $\mathcal{K}_{ii}$.* In the proof of Theorem 5.2, assumption (5.27) is used to conclude that various G_{kl} and T_{kl} with $k, l = i$ or $\hat{i}$ are finite. More specifically, with the aid of assumption (5.27) and with the upper bound of $|\omega_B|$ and the lower bound on $\operatorname{Im}\omega_B$ in (A.4) that together imply that ω_B^2 is away from the positive real axis so the denominators in the definition of $\Lambda_{\mathrm{d},ij}$ do not vanish, we have

$$\max_{i \in [\![1,N]\!]} \max_{k,l=i \text{ or } \hat{i}} |G_{kl}| \prec 1, \qquad \max_{i \in [\![1,N]\!]} \max_{k,l=i \text{ or } \hat{i}} |T_{kl}| \prec 1. \tag{5.31}$$

In addition, using the identities in (4.16), we can further get the bound

$$\max_{i \in [\![1,N]\!]} \max_{k,l=i \text{ or } \hat{i}} \big|(XGY)_{kl}\big| \prec 1, \qquad X, Y = \hat{I} \text{ or } \widetilde{B}. \tag{5.32}$$

Observe that

$$\frac{1}{N} \sum_i \frac{\omega_B(z)}{|\xi_i|^2 - (\omega_B(z))^2} = m_{\mu_A}(\omega_B(z)) = m_{\mu_A \boxplus \mu_B}(z), \tag{5.33}$$

where the first step follows from the definition of μ_A in (4.19), and the second step follows from (2.5) with the choice $(\mu_1, \mu_2) = (\mu_A, \mu_B)$. Then (5.33) together

with the first estimate in (5.27), (4.16) and the upper bound of $|\omega_B|$ and the lower bound of $\operatorname{Im}\omega_B$ in (A.4) leads to the following estimates for tracial quantities:

$$\begin{aligned}
\tau_a(G) &= m_{\mu_A\boxplus\mu_B} + O_{\prec}\big(N^{-\frac{\gamma}{4}}\big), a = 1, 2,\\
\tau_a(\widetilde{B}G) &= (z-\omega_B)m_{\mu_A\boxplus\mu_B} + O_{\prec}\big(N^{-\frac{\gamma}{4}}\big),\\
\tau_a(G\widetilde{B}) &= (z-\omega_B)m_{\mu_A\boxplus\mu_B} + O_{\prec}\big(N^{-\frac{\gamma}{4}}\big),\\
\tau_a(\widetilde{B}G\widetilde{B}) &= (\omega_B-z)\big(1+(\omega_B-z)m_{\mu_A\boxplus\mu_B}\big) + O_{\prec}\big(N^{-\frac{\gamma}{4}}\big).
\end{aligned} \tag{5.34}$$

Then, using the upper bound on $|\omega_B|$ and the lower bound on $\operatorname{Im}\omega_B$ in (A.4), and the second identity in (5.33), we see that all these tracial quantities are stochastically dominated by 1, under assumption (5.27). Recalling Υ_a from (5.20), we thus have under assumption (5.27) that

$$\big|\Upsilon_a(z)\big| \prec 1. \tag{5.35}$$

For (5.28), we only handle the estimate of $\mathcal{P}_{ii}$ and $\mathcal{K}_{ii}$ in detail. The others are similar. It suffices to show the high order moment estimate: for any fixed integer $p \geq 1$, we have

$$\mathbb{E}\big[|\mathcal{P}_{ii}|^{2p}\big] \prec \Psi^{2p}, \qquad \mathbb{E}\big[|\mathcal{K}_{ii}|^{2p}\big] \prec \Psi^{2p}. \tag{5.36}$$

Let us introduce the notation

$$\mathfrak{m}_i(k,l) := \mathcal{P}_{ii}^k\overline{\mathcal{P}_{ii}^l}, \qquad \mathfrak{n}_i(k,l) := \mathcal{K}_{ii}^k\overline{\mathcal{K}_{ii}^l}. \tag{5.37}$$

We will use the following notational conventions in the statement of the recursive moment estimates. The notation $O_{\prec}(\Psi^k)$ for any given positive integer k, represents a generic (possibly) z-dependent random variable $X \equiv X(z)$ that satisfies

$$X \prec \Psi^k, \qquad \mathbb{E}\big[|X|^q\big] \prec \Psi^{qk}, \tag{5.38}$$

for any given positive integer q. In the sequel, we only check the first bound in (5.38) for various X's, then the second bound is valid as well. Indeed, since the X's we will encounter below are analogous to those in [4], we refer to the paragraph below (6.2) of [4] for a general reasoning why the second bound in (5.38) follows from the first one. Additionally, sometimes X will be of the form $1/|g|$ where g is an N-dimensional Gaussian random variable [see, e.g., (5.56)–(5.57)], whose qth moments are also integrable for any fixed q if N is large enough.

The main technical task in the proof of (5.36) is the following recursive moment estimate.

LEMMA 5.3 (Recursive moment estimate for $\mathcal{P}_{ii}$ and $\mathcal{K}_{ii}$). *Suppose the assumptions of Theorem 5.2 hold. For any fixed integer $p \geq 1$, and for any $i \in [\![1, N]\!]$,*

we have

$$\begin{aligned}\mathbb{E}\big[\mathfrak{m}_i(p,p)\big] &= \mathbb{E}\big[O_\prec(\Psi)\mathfrak{m}_i(p-1,p)\big] + \mathbb{E}\big[O_\prec(\Psi^2)\mathfrak{m}_i(p-2,p)\big] \\ &\quad + \mathbb{E}\big[O_\prec(\Psi^2)\mathfrak{m}_i(p-1,p-1)\big], \\ \mathbb{E}\big[\mathfrak{n}_i(p,p)\big] &= \mathbb{E}\big[O_\prec(\Psi)\mathfrak{n}_i(p-1,p)\big] + \mathbb{E}\big[O_\prec(\Psi^2)\mathfrak{n}_i(p-2,p)\big] \\ &\quad + \mathbb{E}\big[O_\prec(\Psi^2)\mathfrak{n}_i(p-1,p-1)\big],\end{aligned} \tag{5.39}$$

where we made the convention $\mathfrak{m}_i(0,0) = \mathfrak{n}_i(0,0) = 1$ *and* $\mathfrak{m}_i(-1,1) = \mathfrak{n}_i(-1,1) = 0$ *if* $p = 1$.

PROOF. According to the decomposition in (5.7), for $i \in [\![1, N]\!]$, we have

$$\begin{aligned}(\widetilde{B}G)_{ii} &= \hat{\boldsymbol{e}}_i^* \mathcal{R}_i \widetilde{B}^{\langle i\rangle} \mathcal{R}_i G \hat{\boldsymbol{e}}_i = -\big((\boldsymbol{h}_i^u)^*, \boldsymbol{0}^*\big)\widetilde{B}^{\langle i\rangle}\mathcal{R}_i G\hat{\boldsymbol{e}}_i \\ &= -(\boldsymbol{k}_i^u)^* \widetilde{B}^{\langle i\rangle}\mathcal{R}_i G\hat{\boldsymbol{e}}_i,\end{aligned} \tag{5.40}$$

where in the second step we used (5.13), and in the last step we used the notation in (5.15). Using (5.40), the definition in (5.3), and also the identity in (5.11), one can check

$$\begin{aligned}(\widetilde{B}G)_{ii} &= -(\boldsymbol{k}_i^u)^*\widetilde{B}^{\langle i\rangle}(\hat{I} - \boldsymbol{r}_i^u(\boldsymbol{r}_i^u)^* \oplus \boldsymbol{r}_i^v(\boldsymbol{r}_i^v)^*)G\hat{\boldsymbol{e}}_i \\ &= -S_{ii} + (\boldsymbol{k}_i^u)^*\widetilde{B}^{\langle i\rangle}(\boldsymbol{r}_i^u(\boldsymbol{r}_i^u)^* \oplus \boldsymbol{r}_i^v(\boldsymbol{r}_i^v)^*)G\hat{\boldsymbol{e}}_i \\ &= -S_{ii} + (\boldsymbol{k}_i^u)^*\widetilde{B}^{\langle i\rangle}(0 \oplus \boldsymbol{r}_i^v(\boldsymbol{r}_i^v)^*)G\hat{\boldsymbol{e}}_i \\ &= -S_{ii} + (\ell_i^v)^2(\boldsymbol{k}_i^u)^*\widetilde{B}^{\langle i\rangle}(\hat{\boldsymbol{e}}_{\hat{i}} + \boldsymbol{k}_i^v)(\hat{\boldsymbol{e}}_{\hat{i}} + \boldsymbol{k}_i^v)^* G\hat{\boldsymbol{e}}_i \\ &= -S_{ii} + (\ell_i^v)^2(\tilde{\sigma}_i h_{ii}^u + (\boldsymbol{k}_i^u)^*\widetilde{B}^{\langle i\rangle}\boldsymbol{k}_i^v)(G_{\hat{i}i} + T_{\hat{i}i}) \\ &=: -\mathring{S}_{ii} + \varepsilon_{i1},\end{aligned} \tag{5.41}$$

where 0 in the third line is the $N \times N$ zero matrix, and

$$\begin{aligned}\varepsilon_{i1} &:= \big(((\ell_i^v)^2 - 1)\tilde{\sigma}_i h_{ii}^u + (\ell_i^v)^2(\boldsymbol{k}_i^u)^*\widetilde{B}^{\langle i\rangle}\boldsymbol{k}_i^v\big)G_{\hat{i}i} \\ &\quad + (\ell_i^v)^2(\tilde{\sigma}_i h_{ii}^u + (\boldsymbol{k}_i^u)^*\widetilde{B}^{\langle i\rangle}\boldsymbol{k}_i^v)T_{\hat{i}i}.\end{aligned} \tag{5.42}$$

In the third step of (5.41), we used the fact $(\boldsymbol{k}_i^u)^*\widetilde{B}^{\langle i\rangle}(\boldsymbol{r}_i^u(\boldsymbol{r}_i^u)^* \oplus 0) = 0$ which follows from the definition of $\boldsymbol{k}_i^u$ and $\widetilde{B}^{\langle i\rangle}$ in (5.15) and (5.8); in the fifth step we used the second identity in (5.18); and in the last step, we used (5.17). We note that

$$|\varepsilon_{i1}| \prec \Psi, \tag{5.43}$$

where we used (5.31) and the large deviation bound (A.1) to show that $(\boldsymbol{k}_i^u)^*\widetilde{B}^{\langle i\rangle} \times \boldsymbol{k}_i^v \prec N^{-1/2}$.

Using integration by parts, we note that

$$\int_{\mathbb{C}} \bar{g} f(g,\bar{g}) \mathrm{e}^{-\frac{|g|^2}{\sigma^2}} \mathrm{d}^2 g = \sigma^2 \int_{\mathbb{C}} \partial_g f(g,\bar{g}) \mathrm{e}^{-\frac{|g|^2}{\sigma^2}} \mathrm{d}^2 g, \tag{5.44}$$

for differentiable functions $f : \mathbb{C}^2 \to \mathbb{C}$ (recall that $\mathrm{d}^2 g$ is the Lebesgue measure on $\mathbb{C}$).

According to the definitions in (5.3), (5.10) and the identity (5.11), one can check for $k \neq i$,

$$\frac{\partial R_i^a}{\partial g_{ik}^a} = -\frac{(\ell_i^a)^2}{\|\boldsymbol{g}_i^a\|_2} \boldsymbol{e}_k (\boldsymbol{e}_i + \boldsymbol{h}_i^a)^* + \Delta_R^a(i,k), \qquad a = u, v. \tag{5.45}$$

where

$$\begin{aligned} \Delta_R^a(i,k) := {} & \frac{(\ell_i^a)^2}{2\|\boldsymbol{g}_i^a\|_2^2} \bar{g}_{ik}^a \big(\boldsymbol{e}_i (\boldsymbol{h}_i^a)^* + \boldsymbol{h}_i^a \boldsymbol{e}_i^* + 2\boldsymbol{h}_i^a (\boldsymbol{h}_i^a)^*\big) \\ & - \frac{(\ell_i^a)^4}{2\|\boldsymbol{g}_i^a\|_2^3} g_{ii}^a \bar{g}_{ik}^a (\boldsymbol{e}_i + \boldsymbol{h}_i^a)(\boldsymbol{e}_i + \boldsymbol{h}_i^a)^*, \qquad a = u, v. \end{aligned} \tag{5.46}$$

The $\Delta_R^a(i,k)$'s are irrelevant error terms. Their estimates will be presented separately in Appendix B of the Supplementary Material [5]. For convenience, we set for $a = u, v$,

$$\begin{aligned} c_i^a := \frac{(\ell_i^a)^2}{\|\boldsymbol{g}_i^a\|_2} &= \frac{1}{\|\boldsymbol{g}_i^a\|_2} - h_{ii}^a + O_{\prec}\Big(\frac{1}{N}\Big) \\ &= \|\boldsymbol{g}_i^a\|_2 - h_{ii}^a - (\|\boldsymbol{g}_i^a\|_2^2 - 1) + O_{\prec}\Big(\frac{1}{N}\Big), \end{aligned} \tag{5.47}$$

where the last step follows from (5.12). Using (5.7), we have for $k \neq i$,

$$\frac{\partial G}{\partial g_{ik}^u} = -G \frac{\partial \widetilde{B}}{\partial g_{ik}^u} G = -G \frac{\partial \mathcal{R}_i}{\partial g_{ik}^u} \widetilde{B}^{\langle i \rangle} \mathcal{R}_i G - G \mathcal{R}_i \widetilde{B}^{\langle i \rangle} \frac{\partial \mathcal{R}_i}{\partial g_{ik}^u} G. \tag{5.48}$$

According to (5.45) and the fact $\mathcal{R}_i = R_i^u \oplus R_i^v$, we have

$$\frac{\partial \mathcal{R}_i}{\partial g_{ik}^u} = -c_i^u \hat{\boldsymbol{e}}_k (\hat{\boldsymbol{e}}_i + \boldsymbol{k}_i^u)^* + \Delta_R^a(i,k) \oplus 0, \tag{5.49}$$

where 0 is the $N \times N$ zero matrix. We also used that $\partial R_i^v / \partial g_{ik}^u = 0$. Plugging (5.49) into (5.48), for $k \neq i$, we can write

$$\begin{aligned} \frac{\partial G}{\partial g_{ik}^u} = {} & c_i^u G \hat{\boldsymbol{e}}_k (\hat{\boldsymbol{e}}_i^* + (\boldsymbol{k}_i^u)^*) \widetilde{B}^{\langle i \rangle} \mathcal{R}_i G \\ & + c_i^u G \mathcal{R}_i \widetilde{B}^{\langle i \rangle} \hat{\boldsymbol{e}}_k (\hat{\boldsymbol{e}}_i^* + (\boldsymbol{k}_i^u)^*) G + \Delta_G^u(i,k), \end{aligned} \tag{5.50}$$

where we set

$$\Delta_G^u(i,k) := -G\big(\Delta_R^u(i,k) \oplus 0\big) \widetilde{B}^{\langle i \rangle} \mathcal{R}_i G - G \mathcal{R}_i \widetilde{B}^{\langle i \rangle} \big(\Delta_R^u(i,k) \oplus 0\big) G. \tag{5.51}$$

With the above derivatives, we are ready to apply the integration by parts formula in (5.44). We start with the following:

$$\begin{aligned}\mathbb{E}\big[\mathfrak{m}_i(p,p)\big] &= \mathbb{E}\big[\mathcal{P}_{ii}\mathfrak{m}_i(p-1,p)\big]\\ &= \mathbb{E}\big[(\widetilde{B}G)_{ii}\tau_1(G)\mathfrak{m}_i(p-1,p)\big]\\ &\quad + \mathbb{E}\big[\big(-G_{ii}\tau_1(\widetilde{B}G) + (G_{ii}+T_{ii})\Upsilon_1\big)\mathfrak{m}_i(p-1,p)\big],\end{aligned} \tag{5.52}$$

$$\begin{aligned}\mathbb{E}\big[\mathfrak{n}_i(p,p)\big] &= \mathbb{E}\big[\mathcal{K}_{ii}\mathfrak{n}_i(p-1,p)\big]\\ &= \mathbb{E}\big[T_{ii}\mathfrak{n}_i(p-1,p)\big]\\ &\quad + \mathbb{E}\big[\big(\tau_1(G)\big(\tilde{\sigma}_i T_{\hat{i}j} + (\widetilde{B}G)_{ij}\big) - \tau_1(G\widetilde{B})(G_{ij}+T_{ij})\big)\\ &\quad \times \mathfrak{n}_i(p-1,p)\big],\end{aligned} \tag{5.53}$$

which follow from the definitions in (5.19) and (5.37) directly. From (5.41) and (5.17), we have

$$\begin{aligned}\mathbb{E}\big[(\widetilde{B}G)_{ii}\tau_1(G)\mathfrak{m}_i(p-1,p)\big] &= -\mathbb{E}\big[\mathring{S}_{ii}\tau_1(G)\mathfrak{m}_i(p-1,p)\big]\\ &\quad + \mathbb{E}\big[\varepsilon_{i1}\tau_1(G)\mathfrak{m}_i(p-1,p)\big],\end{aligned} \tag{5.54}$$

$$\begin{aligned}\mathbb{E}\big[T_{ii}\mathfrak{n}_i(p-1,p)\big] &= \mathbb{E}\big[\mathring{T}_{ii}\mathfrak{n}_i(p-1,p)\big]\\ &\quad + \mathbb{E}\big[O_{\prec}(\Psi)\mathfrak{n}_i(p-1,p)\big],\end{aligned} \tag{5.55}$$

where we used the fact $|h_{ii}| \prec N^{-\frac{1}{2}}$, and also (5.31).

Now we will carefully compute the first terms in the right-hand side of (5.54) and (5.55) with the integration by parts formula since both $\mathring{S}_{ii}$ and $\mathring{T}_{ii}$ explicitly contain a multiplicative Gaussian factor. We will then find that the leading term of the result of this calculation will exactly cancel the last quantities in the right-hand side of equations in (5.52) and (5.53). This cancellation is the key point of the following tedious calculation and this is the main reason for defining the key quantities $\mathcal{P}_{ii}$ and $\mathcal{K}_{ii}$ in the form they are given in (5.19).

For the first term on the right-hand side of (5.54), using the definition of $\mathring{S}_{ii}$ in (5.17) and the integration by parts formula in (5.44), we have

$$\begin{aligned}&\mathbb{E}\big[\mathring{S}_{ii}\tau_1(G)\mathfrak{m}_i(p-1,p)\big]\\ &\quad = \sum_k^{(i)} \mathbb{E}\bigg[\bar{g}^u_{ik}\frac{1}{\|\boldsymbol{g}^u_i\|_2}\hat{\boldsymbol{e}}_k^*\widetilde{B}^{\langle i\rangle}G\hat{\boldsymbol{e}}_i\tau_1(G)\mathfrak{m}_i(p-1,p)\bigg]\\ &\quad = \frac{1}{N}\sum_k^{(i)} \mathbb{E}\bigg[\frac{1}{\|\boldsymbol{g}^u_i\|_2}\frac{\partial(\hat{\boldsymbol{e}}_k^*\widetilde{B}^{\langle i\rangle}G\hat{\boldsymbol{e}}_i)}{\partial g^u_{ik}}\tau_1(G)\mathfrak{m}_i(p-1,p)\bigg]\\ &\quad\quad + \frac{1}{N}\sum_k^{(i)} \mathbb{E}\bigg[\frac{\partial\|\boldsymbol{g}^u_i\|_2^{-1}}{\partial g^u_{ik}}\hat{\boldsymbol{e}}_k^*\widetilde{B}^{\langle i\rangle}G\hat{\boldsymbol{e}}_i\tau_1(G)\mathfrak{m}_i(p-1,p)\bigg]\end{aligned} \tag{5.56}$$

$$+\frac{1}{N}\sum_k^{(i)}\mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i^u\|_2}\hat{\boldsymbol{e}}_k^*\widetilde{B}^{\langle i\rangle}G\hat{\boldsymbol{e}}_i\frac{\partial\tau_1(G)}{\partial g_{ik}^u}\mathfrak{m}_i(p-1,p)\Big]$$

$$+\frac{p-1}{N}\sum_k^{(i)}\mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i^u\|_2}\hat{\boldsymbol{e}}_k^*\widetilde{B}^{\langle i\rangle}G\hat{\boldsymbol{e}}_i\tau_1(G)\frac{\partial\mathcal{P}_{ii}}{\partial g_{ik}^u}\mathfrak{m}_i(p-2,p)\Big]$$

$$+\frac{p}{N}\sum_k^{(i)}\mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i^u\|_2}\hat{\boldsymbol{e}}_k^*\widetilde{B}^{\langle i\rangle}G\hat{\boldsymbol{e}}_i\tau_1(G)\frac{\partial\overline{\mathcal{P}_{ii}}}{\partial g_{ik}^u}\mathfrak{m}_i(p-1,p-1)\Big].$$

Analogously, we have

$$\begin{aligned}
&\mathbb{E}\big[\mathring{T}_{ii}\mathfrak{n}_i(p-1,p)\big]\\
&\quad=\frac{1}{N}\sum_k^{(i)}\mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i^u\|_2}\frac{\partial(\hat{\boldsymbol{e}}_k^*G\hat{\boldsymbol{e}}_i)}{\partial g_{ik}^u}\mathfrak{n}_i(p-1,p)\Big]\\
&\qquad+\frac{1}{N}\sum_k^{(i)}\mathbb{E}\Big[\frac{\partial\|\boldsymbol{g}_i^u\|_2^{-1}}{\partial g_{ik}^u}\hat{\boldsymbol{e}}_k^*G\hat{\boldsymbol{e}}_i\mathfrak{n}_i(p-1,p)\Big]\\
&\qquad+\frac{p-1}{N}\sum_k^{(i)}\mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i^u\|_2}\hat{\boldsymbol{e}}_k^*G\hat{\boldsymbol{e}}_i\frac{\partial\mathcal{K}_{ii}}{\partial g_{ik}^u}\mathfrak{n}_i(p-2,p)\Big]\\
&\qquad+\frac{p}{N}\sum_k^{(i)}\mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i^u\|_2}\hat{\boldsymbol{e}}_k^*G\hat{\boldsymbol{e}}_i\frac{\partial\overline{\mathcal{K}_{ii}}}{\partial g_{ik}^u}\mathfrak{n}_i(p-1,p-1)\Big].
\end{aligned}\tag{5.57}$$

We start from the first term on the right-hand side of (5.56). Using (5.50), we have

$$\begin{aligned}
\frac{1}{N}\sum_k^{(i)}\frac{\partial(\hat{\boldsymbol{e}}_k^*\widetilde{B}^{\langle i\rangle}G\hat{\boldsymbol{e}}_i)}{\partial g_{ik}^u}&=c_i^u\frac{1}{N}\sum_k^{(i)}\hat{\boldsymbol{e}}_k^*\widetilde{B}^{\langle i\rangle}G\hat{\boldsymbol{e}}_k\big(\hat{\boldsymbol{e}}_i+\boldsymbol{k}_i^u\big)^*\widetilde{B}^{\langle i\rangle}\mathcal{R}_iG\hat{\boldsymbol{e}}_i\\
&\quad+c_i^u\frac{1}{N}\sum_k^{(i)}\hat{\boldsymbol{e}}_k^*\widetilde{B}^{\langle i\rangle}G\mathcal{R}_i\widetilde{B}^{\langle i\rangle}\hat{\boldsymbol{e}}_k\big(\hat{\boldsymbol{e}}_i+\boldsymbol{k}_i^u\big)^*G\hat{\boldsymbol{e}}_i\\
&\quad+\frac{1}{N}\sum_k^{(i)}\hat{\boldsymbol{e}}_k^*\widetilde{B}^{\langle i\rangle}\Delta_G^u(i,k)\hat{\boldsymbol{e}}_i.
\end{aligned}\tag{5.58}$$

Let

$$\varepsilon_{i2}:=\frac{1}{N}\sum_k^{(i)}\hat{\boldsymbol{e}}_k^*\widetilde{B}^{\langle i\rangle}\Delta_G^u(i,k)\hat{\boldsymbol{e}}_i.\tag{5.59}$$

Note that

$$\frac{1}{N}\sum_{k}^{(i)}\hat{\boldsymbol{e}}_k^*\widetilde{B}^{\langle i\rangle}G\hat{\boldsymbol{e}}_k=\tau_1\big(\widetilde{B}^{\langle i\rangle}G\big)-\frac{1}{N}\big(\widetilde{B}^{\langle i\rangle}G\big)_{ii}=\tau_1(\widetilde{B}G)+O_{\prec}\big(\Psi^2\big),\tag{5.60}$$

where in the last step we used the second estimate in Corollary A.4 with the choice $Q=\hat{I}_1$ (cf. (4.25)), $(\widetilde{B}^{\langle i\rangle}G)_{ii}=\tilde{\sigma}_i G_{\hat{i}\hat{i}}$ (cf. (5.18)), and the bound in (5.31). Analogously, one shows

$$\frac{1}{N}\sum_{k}^{(i)}\hat{\boldsymbol{e}}_k^*\widetilde{B}^{\langle i\rangle}G\mathcal{R}_i\widetilde{B}^{\langle i\rangle}\hat{\boldsymbol{e}}_k=\tau_1(\widetilde{B}G\widetilde{B})+O_{\prec}\big(\Psi^2\big).\tag{5.61}$$

Moreover, using (5.18), (5.13) and the fact $\mathcal{R}_i^2=\hat{I}$, we also have the following observations:

$$\begin{aligned}\hat{\boldsymbol{e}}_i^*\widetilde{B}^{\langle i\rangle}\mathcal{R}_iG\hat{\boldsymbol{e}}_i&=\tilde{\sigma}_i\hat{\boldsymbol{e}}_{\hat{i}}^*\mathcal{R}_iG\hat{\boldsymbol{e}}_i=-\tilde{\sigma}_i\big(\boldsymbol{k}_i^v\big)^*G\hat{\boldsymbol{e}}_i=-\tilde{\sigma}_iT_{\hat{i}\hat{i}},\\ \big(\boldsymbol{k}_i^u\big)^*\widetilde{B}^{\langle i\rangle}\mathcal{R}_iG\hat{\boldsymbol{e}}_i&=\big(\boldsymbol{k}_i^u\big)^*\mathcal{R}_i\widetilde{B}G\hat{\boldsymbol{e}}_i=-\hat{\boldsymbol{e}}_i^*\widetilde{B}G\hat{\boldsymbol{e}}_i=-(\widetilde{B}G)_{ii}.\end{aligned}\tag{5.62}$$

Plugging (5.60), (5.61) and (5.62) into (5.58), we obtain

$$\begin{aligned}\frac{1}{N}\sum_{k}^{(i)}\frac{\partial(\hat{\boldsymbol{e}}_k^*\widetilde{B}^{\langle i\rangle}G\hat{\boldsymbol{e}}_i)}{\partial g_{ik}^u}&=-c_i^u\tau_1(\widetilde{B}G)\big(\tilde{\sigma}_iT_{\hat{i}\hat{i}}+(\widetilde{B}G)_{ii}\big)\\&\quad+c_i^u\tau_1(\widetilde{B}G\widetilde{B})(G_{ii}+T_{ii})+\varepsilon_{i2}+O_{\prec}\big(\Psi^2\big).\end{aligned}\tag{5.63}$$

Analogously to (5.63), we also have

$$\begin{aligned}\frac{1}{N}\sum_{k}^{(i)}\frac{\partial(\hat{\boldsymbol{e}}_k^*G\hat{\boldsymbol{e}}_i)}{\partial g_{ik}^u}&=-c_i^u\tau_1(G)\big(\tilde{\sigma}_iT_{\hat{i}\hat{i}}+(\widetilde{B}G)_{ii}\big)\\&\quad+c_i^u\tau_1(G\widetilde{B})(G_{ii}+T_{ii})+\varepsilon_{i3}+O_{\prec}\big(\Psi^2\big),\end{aligned}\tag{5.64}$$

where

$$\varepsilon_{i3}:=\frac{1}{N}\sum_{k}^{(i)}\hat{\boldsymbol{e}}_k^*\Delta_G^u(i,k)\hat{\boldsymbol{e}}_i.$$

The following estimates on ε_{i2} and ε_{i3} will be proved in Lemma B.1 in Appendix B of the Supplementary Material [5]:

$$|\varepsilon_{i2}|\prec\Psi^2,\qquad|\varepsilon_{i3}|\prec\Psi^2.\tag{5.65}$$

Combining (5.63), (5.64) with an appropriate linear combination and using (5.65), we get

$$\begin{aligned}&\frac{1}{N}\sum_{k}^{(i)}\frac{\partial(\hat{\boldsymbol{e}}_k^*\widetilde{B}^{\langle i\rangle}G\hat{\boldsymbol{e}}_i)}{\partial g_{ik}^u}\tau_1(G)-\frac{1}{N}\sum_{k}^{(i)}\frac{\partial(\hat{\boldsymbol{e}}_k^*G\hat{\boldsymbol{e}}_i)}{\partial g_{ik}^u}\tau_1(\widetilde{B}G)\\&\quad=-c_i^u(G_{ii}+T_{ii})\big(\tau_1(\widetilde{B}G)-\Upsilon_1\big)+O_{\prec}\big(\Psi^2\big).\end{aligned}\tag{5.66}$$

Here we also used that the tracial quantities $\tau_1(G)$, $\tau_1(\widetilde{B}G)$ and Υ_1 are stochastically dominated by 1, in light of (5.34). Applying (5.47), the fact $\mathring{T}_{ii} = T_{ii} - h_{ii}^u G_{ii}$ from (5.17), we can write

$$\begin{aligned}
&\frac{1}{N}\sum_k^{(i)} \frac{\partial(\hat{\boldsymbol{e}}_k^* \widetilde{B}^{\langle i\rangle} G \hat{\boldsymbol{e}}_i)}{\partial g_{ik}^u} \tau_1(G) \\
&\quad = -c_i^u (G_{ii} + T_{ii})\big(\tau_1(\widetilde{B}G) - \Upsilon_1\big) \\
&\qquad + \frac{1}{N}\sum_k^{(i)} \frac{\partial(\hat{\boldsymbol{e}}_k^* G \hat{\boldsymbol{e}}_i)}{\partial g_{ik}^u} \tau_1(\widetilde{B}G) + O_{\prec}(\Psi^2) \\
&\quad = -c_i^u (G_{ii} + T_{ii})\big(\tau_1(\widetilde{B}G) - \Upsilon_1\big) + \mathring{T}_{ii}\tau_1(\widetilde{B}G) \\
&\qquad + \bigg(\frac{1}{N}\sum_k^{(i)} \frac{\partial(\hat{\boldsymbol{e}}_k^* G \hat{\boldsymbol{e}}_i)}{\partial g_{ik}^u} - \mathring{T}_{ii}\bigg)\tau_1(\widetilde{B}G) + O_{\prec}(\Psi^2) \\
&\quad = -\|\boldsymbol{g}_i^u\|_2 (G_{ii}\tau_1(\widetilde{B}G) - (G_{ii} + T_{ii})\Upsilon_1) \\
&\qquad + \bigg(\frac{1}{N}\sum_k^{(i)} \frac{\partial(\hat{\boldsymbol{e}}_k^* G \hat{\boldsymbol{e}}_i)}{\partial g_{ik}^u} - \mathring{T}_{ii}\bigg)\tau_1(\widetilde{B}G) + \varepsilon_{i4} + \varepsilon_{i5} + O_{\prec}(\Psi^2),
\end{aligned} \tag{5.67}$$

where

$$\begin{aligned}
\varepsilon_{i4} :=\ & \big((1 - \|\boldsymbol{g}_i^u\|_2^2)\tau_1(\widetilde{B}G) + (1 - \|\boldsymbol{g}_i^u\|_2^2 - h_{ii})(\tau_1(\widetilde{B}G) - \Upsilon_1)\big)T_{ii} \\
&+ (1 - \|\boldsymbol{g}_i^u\|_2^2 - h_{ii})G_{ii}\Upsilon_1,
\end{aligned} \tag{5.68}$$

$$\varepsilon_{i5} := (\|\boldsymbol{g}_i^u\|_2^2 - 1)G_{ii}\tau_1(\widetilde{B}G). \tag{5.69}$$

Using $\|\boldsymbol{g}_i^u\|_2 = 1 + O_{\prec}(\frac{1}{\sqrt{N}})$, the estimates (5.31), (5.32) and (5.34), and Corollary A.4, we get

$$|\varepsilon_{i4}| \prec \frac{1}{\sqrt{N}}, \qquad |\varepsilon_{i5}| \prec \frac{1}{\sqrt{N}}. \tag{5.70}$$

Notice that the first term in the right-hand side of (5.67) will exactly cancel the explicit last term in the right-hand side of (5.52). This cancellation is one of the main reasons behind the choice of the auxiliary quantity $\mathcal{P}$. Combining the first equation of (5.53), (5.54), (5.56) with (5.67), we get

$$\begin{aligned}
\mathbb{E}\big[\mathfrak{m}_i(p,p)\big] =\ & \mathbb{E}\bigg[\frac{1}{\|\boldsymbol{g}_i^u\|_2}\bigg(\mathring{T}_{ii} - \frac{1}{N}\sum_k^{(i)} \frac{\partial(\hat{\boldsymbol{e}}_k^* G \hat{\boldsymbol{e}}_i)}{\partial g_{ik}^u}\bigg)\tau_1(\widetilde{B}G)\mathfrak{m}_i(p-1,p)\bigg] \\
&- \frac{1}{N}\sum_k^{(i)} \mathbb{E}\bigg[\frac{\partial \|\boldsymbol{g}_i^u\|_2^{-1}}{\partial g_{ik}^u} \hat{\boldsymbol{e}}_k^* \widetilde{B}^{\langle i\rangle} G \hat{\boldsymbol{e}}_i \tau_1(G)\mathfrak{m}_i(p-1,p)\bigg]
\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{N}\sum_k^{(i)} \mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i^u\|_2}\hat{\boldsymbol{e}}_k^* \widetilde{B}^{\langle i\rangle} G\hat{\boldsymbol{e}}_i \frac{\partial \tau_1(G)}{\partial g_{ik}^u}\mathfrak{m}_i(p-1,p)\Big] \\
&- \frac{p-1}{N}\sum_k^{(i)} \mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i^u\|_2}\hat{\boldsymbol{e}}_k^* \widetilde{B}^{\langle i\rangle} G\hat{\boldsymbol{e}}_i \tau_1(G)\frac{\partial \mathcal{P}_{ii}}{\partial g_{ik}^u}\mathfrak{m}_i(p-2,p)\Big] \\
&- \frac{p}{N}\sum_k^{(i)} \mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i^u\|_2}\hat{\boldsymbol{e}}_k^* \widetilde{B}^{\langle i\rangle} G\hat{\boldsymbol{e}}_i \tau_1(G)\frac{\partial \overline{\mathcal{P}_{ii}}}{\partial g_{ik}^u}\mathfrak{m}_i(p-1,p-1)\Big] \\
&+ \mathbb{E}\Big[\Big(\varepsilon_{i1}\tau_1(G) - \frac{\varepsilon_{i4}+\varepsilon_{i5}}{\|\boldsymbol{g}_i^u\|_2}\Big)\mathfrak{m}_i(p-1,p)\Big] \\
&+ \mathbb{E}\big[O_\prec(\Psi^2)\mathfrak{m}_i(p-1,p)\big].
\end{aligned} \tag{5.71}
$$

Note that the sixth term on the right-hand side can be estimated by $\mathbb{E}[O_\prec(\Psi)\times \mathfrak{m}_i(p-1,p)]$, according to (5.43) and (5.70). This estimate is sufficient for the proof of Lemma 5.3. But here we keep the ε-terms explicit for further use.

In order to estimate the first term in the right-hand side, similar to (5.57), we can apply the integration by parts formula (5.44) to obtain

$$
\begin{aligned}
&\mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i^u\|_2}\Big(\mathring{T}_{ii} - \frac{1}{N}\sum_k^{(i)}\frac{\partial(\hat{\boldsymbol{e}}_k^* G\hat{\boldsymbol{e}}_i)}{\partial g_{ik}^u}\Big)\tau_1(\widetilde{B}G)\mathfrak{m}_i(p-1,p)\Big] \\
&\quad= \frac{1}{N}\sum_k^{(i)}\mathbb{E}\Big[\frac{\partial\|\boldsymbol{g}_i^u\|_2^{-2}}{\partial g_{ik}^u}\hat{\boldsymbol{e}}_k^* G\hat{\boldsymbol{e}}_i\tau_1(\widetilde{B}G)\mathfrak{m}_i(p-1,p)\Big] \\
&\quad\quad+ \frac{p-1}{N}\sum_k^{(i)}\mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i^u\|_2^2}\hat{\boldsymbol{e}}_k^* G\hat{\boldsymbol{e}}_i\frac{\partial\tau_1(\widetilde{B}G)}{\partial g_{ik}^u}\mathfrak{m}_i(p-1,p)\Big] \\
&\quad\quad+ \frac{p-1}{N}\sum_k^{(i)}\mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i^u\|_2^2}\hat{\boldsymbol{e}}_k^* G\hat{\boldsymbol{e}}_i\tau_1(\widetilde{B}G)\frac{\partial\mathcal{P}_{ii}}{\partial g_{ik}^u}\mathfrak{m}_i(p-2,p)\Big] \\
&\quad\quad+ \frac{p}{N}\sum_k^{(i)}\mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i^u\|_2^2}\hat{\boldsymbol{e}}_k^* G\hat{\boldsymbol{e}}_i\tau_1(\widetilde{B}G)\frac{\partial\overline{\mathcal{P}_{ii}}}{\partial g_{ik}^u}\mathfrak{m}_i(p-1,p-1)\Big].
\end{aligned} \tag{5.72}
$$

Notice the cancellation between the two terms in the bracket in the first line.

Next we consider the estimate of $\mathfrak{n}_i(p,p)$; especially we control the first term in the right-hand side of (5.57). In addition, using (5.64), (5.65) and the facts $\|\boldsymbol{g}_i^u\|_2 = 1+O_\prec(\frac{1}{\sqrt{N}})$ and $c_i^u = 1+O_\prec(\frac{1}{\sqrt{N}})$, we have

$$
\begin{aligned}
&\frac{1}{N}\frac{1}{\|\boldsymbol{g}_i^u\|_2}\sum_k^{(i)}\frac{\partial(\hat{\boldsymbol{e}}_k^* G\hat{\boldsymbol{e}}_i)}{\partial g_{ik}^u} \\
&\quad= -\tau_1(G)\big(\tilde{\sigma}_i T_{\hat{i}i} + (\widetilde{B}G)_{ii}\big) + \tau_1(G\widetilde{B})(G_{ii}+T_{ii}) + O_\prec(\Psi).
\end{aligned} \tag{5.73}
$$

Note that the result of this calculation exactly cancels the second term in the right-hand side of (5.53). Hence, analogously to (5.71), combining (5.57), (5.65), (5.55), (5.53) and (5.73), we get

$$\begin{aligned}\mathbb{E}\big[\mathfrak{n}_i(p,p)\big] &= \frac{1}{N}\sum_k^{(i)}\mathbb{E}\bigg[\frac{\partial\|\boldsymbol{g}_i^u\|_2^{-1}}{\partial g_{ik}^u}\hat{\boldsymbol{e}}_k^* G\hat{\boldsymbol{e}}_i\,\mathfrak{n}_i(p-1,p)\bigg]\\ &\quad+\frac{p-1}{N}\sum_k^{(i)}\mathbb{E}\bigg[\frac{1}{\|\boldsymbol{g}_i^u\|_2}\hat{\boldsymbol{e}}_k^* G\hat{\boldsymbol{e}}_i\frac{\partial\mathcal{K}_{ii}}{\partial g_{ik}^u}\mathfrak{n}_i(p-2,p)\bigg]\\ &\quad+\frac{p}{N}\sum_k^{(i)}\mathbb{E}\bigg[\frac{1}{\|\boldsymbol{g}_i^u\|_2}\hat{\boldsymbol{e}}_k^* G\hat{\boldsymbol{e}}_i\frac{\partial\overline{\mathcal{K}_{ii}}}{\partial g_{ik}^u}\mathfrak{n}_i(p-1,p-1)\bigg]\\ &\quad+\mathbb{E}\big[O_\prec(\Psi)\mathfrak{n}_i(p-1,p)\big].\end{aligned}\tag{5.74}$$

Hence, to prove the second equation of (5.39) it suffices to estimate the first three terms on the right-hand side of (5.74). For the first equation of (5.39), with (5.43) and (5.70), it suffices to estimate the second to the fifth terms on the right-hand side of (5.71), and the terms on the right-hand side of (5.72). All these estimates can be derived from the following lemma.

LEMMA 5.4. *Suppose that the assumptions in Theorem 5.2 hold. Set $X_i=\hat{I}$ or $\widetilde{B}^{\langle i\rangle}$. Let Q be any deterministic diagonal matrix satisfying $\|Q\|\le C$ and $X=\hat{I}$ or A. We have the following estimates*:

$$\begin{aligned}\frac{1}{N}\sum_k^{(i)}\frac{\partial\|\boldsymbol{g}_i^u\|_2^{-1}}{\partial g_{ik}^u}\hat{\boldsymbol{e}}_k^* X_i G\hat{\boldsymbol{e}}_i &= O_\prec\Big(\frac{1}{N}\Big),\\ \frac{1}{N}\sum_k^{(i)}\hat{\boldsymbol{e}}_i^* X\frac{\partial G}{\partial g_{ik}^u}\hat{\boldsymbol{e}}_i\hat{\boldsymbol{e}}_k^* X_i G\hat{\boldsymbol{e}}_i &= O_\prec(\Psi^2),\\ \frac{1}{N}\sum_k^{(i)}\frac{\partial T_{ji}}{\partial g_{ik}^u}\hat{\boldsymbol{e}}_k^* X_i G\hat{\boldsymbol{e}}_i &= O_\prec(\Psi^2),\\ \frac{1}{N}\sum_k^{(i)}\frac{\partial\operatorname{tr}QXG}{\partial g_{ik}^u}\hat{\boldsymbol{e}}_k^* X_i G\hat{\boldsymbol{e}}_i &= O_\prec(\Psi^4),\end{aligned}\tag{5.75}$$

where $j=i$ or $\hat{\imath}$ in the third equation.

Assuming the validity of Lemma 5.4, we continue with the proof of Lemma 5.3. Recall that our task is to bound the terms on the right-hand sides of (5.71), (5.72), (5.74). The second term in (5.71), the first term in (5.72) and the first term in (5.74) can all be estimated with the aid of first bound in (5.75). The estimates for

the third term in (5.71) and the second term in (5.72) follow from the last bound in (5.75). Finally, the fourth term in (5.71), the third term in (5.72) and the second term in (5.74) together with their complex conjugate analogues can be estimated in a similar way, so we only present the details for the fourth term on the right-hand side of (5.71) in the sequel.

Recall the definition of $\mathcal{P}_{ii}$ from (5.19),

$$\mathcal{P}_{ii} = (\widetilde{B}G)_{ii}\tau_1(G) - G_{ii}\tau_1(\widetilde{B}G) + (G_{ii} + T_{ii})\Upsilon_1.$$

Using (4.16), and recalling the definition of Υ_1 in (5.20), we can see that $\mathcal{P}_{ii}$ is a combination of the terms of the following forms: T_{ii}, $(XG)_{ii}$ and $\mathrm{tr}(QXG)$, for $X = \hat{I}$ or A, and Q is certain deterministic diagonal matrix with $\|Q\| \le C$ for some positive constant C. For example, $(\widetilde{B}G)_{ii} = 1 + zG_{ii} - (AG)_{ii}$, and

$$\begin{aligned}\tau_1(G\widetilde{B}) &= \tau_1\big(\hat{I} - G(A - z)\big) = 1 + z\tau_1(G) - \tau_1(GA)\\ &= 1 + 2z\,\mathrm{tr}(\hat{I}_1 G) - 2\,\mathrm{tr}(A\hat{I}_1 G) = 1 + 2z\,\mathrm{tr}(\hat{I}_1 G) - 2\,\mathrm{tr}(\hat{I}_2 AG).\end{aligned}$$

Then, by the product rule for derivative, and the boundedness of all the partial traces (cf. (5.34)) and entries (cf. (5.31), (5.32)), we can apply the last three bounds in (5.75) to conclude that the fourth term on the right-hand side of (5.71) is $\mathbb{E}[O_{\prec}(\Psi^2)\mathfrak{m}_i(p-2,p)]$.

This completes the proof of Lemma 5.3, up to Lemma 5.4. □

PROOF OF LEMMA 5.4. Since the sums in (5.75) are over $k \neq i$, it will be convenient to work in this proof with the following notation:

$$I^{\langle i\rangle} := I - \boldsymbol{e}_i\boldsymbol{e}_i^*, \qquad \hat{I}_1^{\langle i\rangle} := I^{\langle i\rangle} \oplus 0, \tag{5.76}$$

where 0 is the $N \times N$ zero matrix. We check the estimates in (5.75) one by one. For the first estimate, we have

$$\begin{aligned}\frac{1}{N}\sum_k^{(i)} \frac{\partial \|\boldsymbol{g}_i^u\|_2^{-1}}{\partial g_{ik}^u}\hat{\boldsymbol{e}}_k^* X_i G\hat{\boldsymbol{e}}_i &= -\frac{1}{2N}\frac{1}{\|\boldsymbol{g}_i^u\|_2^3}\sum_k^{(i)} \bar{g}_{ik}^u \hat{\boldsymbol{e}}_k^* X_i G\hat{\boldsymbol{e}}_i\\ &= -\frac{1}{2N}\frac{1}{\|\boldsymbol{g}_i^u\|_2^2}(\mathring{\boldsymbol{k}}_i^u)^* X_i G\hat{\boldsymbol{e}}_i = O_{\prec}\Big(\frac{1}{N}\Big),\end{aligned}$$

where in the last step we used that

$$(\mathring{\boldsymbol{k}}_i^u)^* X_i G\hat{\boldsymbol{e}}_i \prec 1, \tag{5.77}$$

which would follow once we show $|\mathring{S}_{ii}| \prec 1$ and $|\mathring{T}_{ii}| = |T_{ii} - h_{ii}^u G_{ii}| \prec 1$ by (5.17). Since $\mathring{S}_{ii} = -(\widetilde{B}G)_{ii} + O_{\prec}(\Psi)$ by (5.41), (5.43) and $|(\widetilde{B}G)_{ii}| \prec 1$ from (5.32), we get $|\mathring{S}_{ii}| \prec 1$. The estimate $|\mathring{T}_{ii}| \prec 1$ follows from (5.31) and the fact $|h_{ii}^u| \prec \frac{1}{\sqrt{N}}$.

Next we show the second estimate in (5.75). Using (5.50), we have

$$
\begin{aligned}
&\frac{1}{N}\sum_{k}^{(i)} \hat{\boldsymbol{e}}_i^* X \frac{\partial G}{\partial g_{ik}^u} \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_k^* X_i G \hat{\boldsymbol{e}}_i \\
&\quad = c_i^u \frac{1}{N}\sum_{k}^{(i)} \hat{\boldsymbol{e}}_i^* X G \hat{\boldsymbol{e}}_k (\hat{\boldsymbol{e}}_i + \boldsymbol{k}_i^u)^* \widetilde{B}^{\langle i\rangle} \mathcal{R}_i G \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_k^* X_i G \hat{\boldsymbol{e}}_i \\
&\qquad + c_i^u \frac{1}{N}\sum_{k}^{(i)} \hat{\boldsymbol{e}}_i^* X G \mathcal{R}_i \widetilde{B}^{\langle i\rangle} \hat{\boldsymbol{e}}_k (\hat{\boldsymbol{e}}_i^* + (\boldsymbol{k}_i^u)^*) G \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_k^* X_i G \hat{\boldsymbol{e}}_i \\
&\qquad + \frac{1}{N}\sum_{k}^{(i)} \hat{\boldsymbol{e}}_i^* X \Delta_G^u(i,k) \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_k^* X_i G \hat{\boldsymbol{e}}_i \\
&\quad = c_i^u \frac{1}{N} \hat{\boldsymbol{e}}_i^* X G \hat{I}_1^{\langle i\rangle} X_i G \hat{\boldsymbol{e}}_i (\hat{\boldsymbol{e}}_i + \boldsymbol{k}_i^u)^* \widetilde{B}^{\langle i\rangle} \mathcal{R}_i G \hat{\boldsymbol{e}}_i \\
&\qquad + c_i^u \frac{1}{N} \hat{\boldsymbol{e}}_i^* X G \mathcal{R}_i \widetilde{B}^{\langle i\rangle} \hat{I}_1^{\langle i\rangle} X_i G \hat{\boldsymbol{e}}_i (\hat{\boldsymbol{e}}_i + \boldsymbol{k}_i^u)^* G \hat{\boldsymbol{e}}_i \\
&\qquad + \frac{1}{N}\sum_{k}^{(i)} \hat{\boldsymbol{e}}_i^* X \Delta_G^u(i,k) \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_k^* X_i G \hat{\boldsymbol{e}}_i,
\end{aligned}
\tag{5.78}
$$

where we have used the notation introduced in (5.76).

From Lemma B.1 in Appendix B of the Supplementary Material [5] we see that the last term on the right-hand side of (5.78) is of order $O_{\prec}(\Psi^2)$. For the first two terms, we first claim that

$$
|\hat{\boldsymbol{e}}_i^* X G \hat{I}_1^{\langle i\rangle} X_i G \hat{\boldsymbol{e}}_i| \prec \frac{1}{\eta}, \qquad |\hat{\boldsymbol{e}}_i^* X G \mathcal{R}_i \widetilde{B}^{\langle i\rangle} \hat{I}_1^{\langle i\rangle} X_i G \hat{\boldsymbol{e}}_i| \prec \frac{1}{\eta}. \tag{5.79}
$$

We prove the first estimate (5.79) as follows. Note that

$$
\begin{aligned}
\hat{\boldsymbol{e}}_i^* X G \hat{I}_i^{\langle i\rangle} X_i G \hat{\boldsymbol{e}}_i &\le \hat{\boldsymbol{e}}_i^* X |G|^2 X \hat{\boldsymbol{e}}_i + \hat{\boldsymbol{e}}_i^* G^* X_i^* \hat{I}_1^{\langle i\rangle} X_i G \hat{\boldsymbol{e}}_i \\
&\le \frac{1}{\eta} \operatorname{Im}(XGX)_{ii} + \|X_i\|^2 \frac{1}{\eta} \operatorname{Im} G_{ii}.
\end{aligned}
\tag{5.80}
$$

Recall $X = \hat{I}$ or A, and the fact $(AGA)_{ii} = |\sigma_i|^2 G_{\hat{i}\hat{i}}$. This together with (5.31) and the fact $\|X_i\| \le C$ since $X_i = \hat{I}$ or $\widetilde{B}^{\langle i\rangle}$ implies the first estimate in (5.79). The second estimate can be derived in a similar way.

Then we recall from (5.62) that $(\hat{\boldsymbol{e}}_i + \boldsymbol{k}_i^u)^* \widetilde{B}^{\langle i\rangle} \mathcal{R}_i G \hat{\boldsymbol{e}}_i = -\widetilde{\sigma}_i T_{\hat{i}i} - (\widetilde{B}G)_{ii}$, and from the definition of T_{ij} in (5.16) that $(\hat{\boldsymbol{e}}_i + \boldsymbol{k}_i^u)^* G \hat{\boldsymbol{e}}_i = G_{ii} + T_{ii}$, which together with (5.31), (5.32) and (5.79) imply that the first two terms on the right-hand side of (5.78) are also of order $O_{\prec}(\Psi^2)$. This completes the second estimate in (5.75).

For the third estimate in (5.75) we present the details for $j = i$ in the sequel. The case of $j = \hat{i}$ is similar but simpler and we omit it. According to the definition of T_{ii} in (5.16), it suffices to show

$$\begin{aligned} &\frac{1}{N}\sum_k^{(i)} \frac{\partial (\boldsymbol{k}_i^u)^*}{\partial g_{ik}^u} G\hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_k^* X_i G\hat{\boldsymbol{e}}_i = O_{\prec}\Big(\frac{1}{N}\Big), \\ &\frac{1}{N}\sum_k^{(i)} (\boldsymbol{k}_i^u)^* \frac{\partial G}{\partial g_{ik}^u} \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_k^* X_i G\hat{\boldsymbol{e}}_i = O_{\prec}(\Psi^2). \end{aligned} \tag{5.81}$$

For the first estimate in (5.81) we have

$$\begin{aligned} \frac{1}{N}\sum_k^{(i)} \frac{\partial (\boldsymbol{k}_i^u)^*}{\partial g_{ik}^u} G\hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_k^* X_i G\hat{\boldsymbol{e}}_i &= -\frac{1}{2\|\boldsymbol{g}_i^u\|_2^2}\frac{1}{N}\sum_k^{(i)} \bar{h}_{ik}^u \hat{\boldsymbol{e}}_k^* X_i G\hat{\boldsymbol{e}}_i (\boldsymbol{k}_i^u)^* G\hat{\boldsymbol{e}}_i \\ &= -\frac{1}{2\|\boldsymbol{g}_i^u\|_2^2}\frac{1}{N}(\mathring{\boldsymbol{k}}_i^u)^* X_i G\hat{\boldsymbol{e}}_i (\boldsymbol{k}_i^u)^* G\hat{\boldsymbol{e}}_i = O_{\prec}\Big(\frac{1}{N}\Big), \end{aligned}$$

where in the last step we used (5.31) and (5.77). The proof of the second estimate in (5.81) is similar to that for the second estimate in (5.75). It suffices to go through the discussion from (5.78) to (5.80) again, with the vector $\hat{\boldsymbol{e}}_i^* X$ replaced by $(\boldsymbol{k}_i^u)^*$. The main differences are: instead of the last term of (5.78), we have

$$\frac{1}{N}\sum_k^{(i)} (\boldsymbol{k}_i^u)^* \Delta_G^u(i,k)\hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_k^* X_i G\hat{\boldsymbol{e}}_i, \tag{5.82}$$

and instead of the first term on the right-hand side of (5.80), we have

$$\frac{1}{\eta}\operatorname{Im}(\boldsymbol{k}_i^u)^* G\boldsymbol{k}_i^u. \tag{5.83}$$

The bound on (5.82) is stated in (B.3). For (5.83), we recall the identity (5.13) which implies $\boldsymbol{k}_i^u = -\mathcal{R}_i\hat{\boldsymbol{e}}_i$, the fact $G = \mathcal{U}\mathcal{G}\mathcal{U}^*$, together with (5.6) and the fact $\mathcal{R}_i^2 = \hat{I}$. Then we have

$$(\boldsymbol{k}_i^u)^* G\boldsymbol{k}_i^u = \hat{\boldsymbol{e}}_i^* \mathcal{R}_i \mathcal{U}\mathcal{G}\mathcal{U}^* \mathcal{R}_i \hat{\boldsymbol{e}}_i = \hat{\boldsymbol{e}}_i^* \mathcal{U}_i \Phi_i \mathcal{G} \Phi_i^* \mathcal{U}_i^* \hat{\boldsymbol{e}}_i = \mathcal{G}_{ii}. \tag{5.84}$$

Similar to (5.31), with the second bound in assumption (5.27), we can also show that

$$\max_{k,l} |\mathcal{G}_{kl}| \prec 1. \tag{5.85}$$

With these bounds for (5.82) and (5.83), we can show the second estimate of (5.81), which together with the first estimate in (5.81) implies the third bound in (5.75).

At the end, we show the last bound in (5.75). Applying (5.50) we have

$$\begin{aligned}\frac{\partial \operatorname{tr} QXG}{\partial g^u_{ik}} &= \frac{1}{N} c_i^u (\hat{\boldsymbol{e}}_i + \boldsymbol{k}_i^u)^* \widetilde{B}^{\langle i\rangle} \mathcal{R}_i G Q X G \hat{\boldsymbol{e}}_k \\ &\quad + \frac{1}{N} c_i^u (\hat{\boldsymbol{e}}_i + \boldsymbol{k}_i^u)^* G Q X G \mathcal{R}_i \widetilde{B}^{\langle i\rangle} \hat{\boldsymbol{e}}_k + \operatorname{tr} Q X \Delta_G^u(i,k).\end{aligned} \tag{5.86}$$

Summing over k and using the notation in (5.76) we can write

$$\begin{aligned}\frac{1}{N}\sum_k^{(i)} \frac{\partial \operatorname{tr} QXG}{\partial g^u_{ik}} \hat{\boldsymbol{e}}_k^* X_i G \hat{\boldsymbol{e}}_i &= \frac{c_i^u}{N^2} (\hat{\boldsymbol{e}}_i + \boldsymbol{k}_i^u)^* \widetilde{B}^{\langle i\rangle} \mathcal{R}_i G Q X G \hat{I}_1^{\langle i\rangle} X_i G \hat{\boldsymbol{e}}_i \\ &\quad + \frac{c_i^u}{N^2} (\hat{\boldsymbol{e}}_i + \boldsymbol{k}_i^u)^* G Q X G \mathcal{R}_i \widetilde{B}^{\langle i\rangle} \hat{I}_1^{\langle i\rangle} X_i G \hat{\boldsymbol{e}}_i \\ &\quad + \frac{1}{N}\sum_k^{(i)} \operatorname{tr} Q X \Delta_G^u(i,k) \hat{\boldsymbol{e}}_k^* X_i G \hat{\boldsymbol{e}}_i.\end{aligned} \tag{5.87}$$

The bound for the last term of the right-hand side of (5.87) can be found in (B.4).

In the sequel, we bound the first two terms on the right-hand side of (5.87). We only present the details for the first one; the second is estimated analogously. First, similar to (5.62), we have

$$(\hat{\boldsymbol{e}}_i + \boldsymbol{k}_i^u)^* \widetilde{B}^{\langle i\rangle} \mathcal{R}_i = -(\tilde{\sigma}_i (\boldsymbol{k}_i^v)^* + \hat{\boldsymbol{e}}_i^* \widetilde{B}).$$

Then we can write

$$\begin{aligned}&\frac{c_i^u}{N^2} (\hat{\boldsymbol{e}}_i + \boldsymbol{k}_i^u)^* \widetilde{B}^{\langle i\rangle} \mathcal{R}_i G Q X G \hat{I}_1^{\langle i\rangle} X_i G \hat{\boldsymbol{e}}_i \\ &\quad = -\frac{c_i^u}{N^2} (\tilde{\sigma}_i (\boldsymbol{k}_i^v)^* + \hat{\boldsymbol{e}}_i^* \widetilde{B}) G Q X G \hat{I}_1 X_i G \hat{\boldsymbol{e}}_i \\ &\qquad + \frac{c_i^u}{N^2} (\tilde{\sigma}_i (\boldsymbol{k}_i^v)^* + \hat{\boldsymbol{e}}_i^* \widetilde{B}) G Q X G \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i^* X_i G \hat{\boldsymbol{e}}_i.\end{aligned} \tag{5.88}$$

For the second term on the right-hand side of (5.88), we use the bounds

$$\big|(\tilde{\sigma}_i (\boldsymbol{k}_i^v)^* + \hat{\boldsymbol{e}}_i^* \widetilde{B}) G Q X G \hat{\boldsymbol{e}}_i\big| \prec \eta^{-2}, \qquad \big|\hat{\boldsymbol{e}}_i^* X_i G \hat{\boldsymbol{e}}_i\big| \prec 1, \tag{5.89}$$

where in the first inequality we used the trivial bound $\|G\| \le \eta^{-1}$, while in the second inequality we used the fact that $X_i = \hat{I}$ or $\widetilde{B}^{\langle i\rangle}$, together with (5.18), and the first bound in (5.31). Using the bounds in (5.89), we see that the second term on the right-hand side of (5.88) is of order $O_\prec(\Psi^4)$.

Now we turn to the first term on the right-hand side of (5.88). Note that

$$\begin{aligned}\left|\frac{1}{N^2}(\tilde{\sigma}_i(\boldsymbol{k}_i^v)^* + \hat{\boldsymbol{e}}_i^*\widetilde{B})GQXG\hat{I}_1X_iG\hat{\boldsymbol{e}}_i\right| &\le \frac{C}{N^2\eta}(\|(\boldsymbol{k}_i^v)^*G\|_2 + \|\hat{\boldsymbol{e}}_i^*\widetilde{B}G\|_2)\|G\hat{\boldsymbol{e}}_i\|_2 \\ &\le \frac{C}{N^2\eta}(\|(\boldsymbol{k}_i^v)^*G\|_2^2 + \|\hat{\boldsymbol{e}}_i^*\widetilde{B}G\|_2^2 + \|G\hat{\boldsymbol{e}}_i\|_2^2) \\ &\le \frac{C}{N^2\eta^2}(\operatorname{Im}(\boldsymbol{k}_i^v)^*G\boldsymbol{k}_i^v + \operatorname{Im}\hat{\boldsymbol{e}}_i^*\widetilde{B}G\widetilde{B}\hat{\boldsymbol{e}}_i + \operatorname{Im}\hat{\boldsymbol{e}}_i^*G\hat{\boldsymbol{e}}_i).\end{aligned} \tag{5.90}$$

Similar to (5.84), we have

$$(\boldsymbol{k}_i^v)^*G\boldsymbol{k}_i^v = \hat{\boldsymbol{e}}_{\hat{i}}^*\mathcal{R}_i\mathcal{U}\mathcal{G}\mathcal{U}^*\mathcal{R}_i\hat{\boldsymbol{e}}_{\hat{i}} = \hat{\boldsymbol{e}}_{\hat{i}}^*\mathcal{U}_i\Phi_i\mathcal{G}\Phi_i^*\mathcal{U}_i^*\hat{\boldsymbol{e}}_{\hat{i}} = \mathcal{G}_{\hat{i}\hat{i}}. \tag{5.91}$$

Combining (5.90) and (5.91), we obtain

$$\begin{aligned}\left|\frac{1}{N^2}(\tilde{\sigma}_i(\boldsymbol{k}_i^v)^* + \hat{\boldsymbol{e}}_i^*\widetilde{B})GQXG\hat{I}_1X_iG\hat{\boldsymbol{e}}_i\right| &\le \frac{C}{N^2\eta^2}(\operatorname{Im}\mathcal{G}_{\hat{i}\hat{i}} + \operatorname{Im}(\widetilde{B}G\widetilde{B})_{ii} + \operatorname{Im}G_{ii}) \\ &= O_{\prec}(\Psi^4),\end{aligned}$$

where we also used (5.32) and (5.85). Hence the first term on the right-hand side of (5.87) is $O_{\prec}(\Psi^4)$. The second term on the right-hand side of (5.87) is bounded similarly. These bounds together with (B.4) yield the other estimates in (5.75). This completes the proof of Lemma 5.4. □

5.4. *Local stability analysis*: *Proof of Theorem* 5.2. Having established Lemma 5.3, we move on to the local stability analysis in order to conclude the proof of Theorem 5.2.

PROOF OF THEOREM 5.2. Applying Young's inequality, we obtain from (5.39) that for any given (small) $\varepsilon > 0$,

$$\mathbb{E}[\mathfrak{m}_i(p,p)] \le 3\frac{1}{2p}\mathbb{E}[N^{2p\varepsilon}\Psi^{2p}] + 3\frac{2p-1}{2p}N^{-\frac{2p\varepsilon}{2p-1}}\mathbb{E}[\mathfrak{m}_i(p,p)],$$

which implies $\mathbb{E}[\mathfrak{m}_i(p,p)] \prec \Psi^{2p}$. Hence, we conclude the proof of the first estimate of (5.36).

The second estimate of (5.36) can be proved in the same way, with the aid of the second equation in (5.39). Then, applying Markov's inequality we get the first and the third estimates of (5.28) with $j = i$. The others in (5.28) are proved in an analogous way. We omit the details.

Next we show that (5.28) together with the assumption (5.27) imply (5.29). To this end, we first show the following crude bound:

$$\Lambda_T(z) \prec N^{-\frac{\gamma}{4}} \tag{5.92}$$

under the assumption (5.27). We need the following equations for $j = i, \hat{i}$:

$$\begin{aligned} T_{ij} &= -\tau_1(G)\big(\tilde{\sigma}_i T_{\hat{i}j} + (\widetilde{B}G)_{ij}\big) + \tau_1(G\widetilde{B})(G_{ij} + T_{ij}) + O_{\prec}(\Psi), \\ T_{\hat{i}j} &= -\tau_2(G)\big(\tilde{\sigma}_i^* T_{ij} + (\widetilde{B}G)_{\hat{i}j}\big) + \tau_2(G\widetilde{B})(G_{\hat{i}j} + T_{\hat{i}j}) + O_{\prec}(\Psi), \end{aligned} \tag{5.93}$$

which is just a rewriting of the second line of (5.28), according to the definition in (5.19).

Using the first identity in (4.16) and the definition of A in (4.14), we have

$$\begin{aligned} (\widetilde{B}G)_{ii} &= 1 + zG_{ii} - \xi_i G_{\hat{i}i}, \qquad (\widetilde{B}G)_{i\hat{i}} = -\xi_i G_{\hat{i}\hat{i}} + zG_{i\hat{i}}, \\ (\widetilde{B}G)_{\hat{i}i} &= -\bar{\xi}_i G_{ii} + zG_{\hat{i}i}, \qquad (\widetilde{B}G)_{\hat{i}\hat{i}} = 1 + zG_{\hat{i}\hat{i}} - \bar{\xi}_i G_{i\hat{i}}. \end{aligned} \tag{5.94}$$

Applying the assumption on Λ_{d} in (5.27), and also the lower bound of $\operatorname{Im}\omega_B$ and the upper bound on $|\omega_B|$ in (A.4), we can get from (5.94) that

$$\begin{aligned} (\widetilde{B}G)_{ii} &= \frac{(z-\omega_B)\omega_B}{|\xi_i|^2 - \omega_B^2} + O_{\prec}(N^{-\frac{\gamma}{4}}), \\ (\widetilde{B}G)_{i\hat{i}} &= \frac{(z-\omega_B)\xi_i}{|\xi_i|^2 - \omega_B^2} + O_{\prec}(N^{-\frac{\gamma}{4}}), \\ (\widetilde{B}G)_{\hat{i}i} &= \frac{(z-\omega_B)\bar{\xi}_i}{|\xi_i|^2 - \omega_B^2} + O_{\prec}(N^{-\frac{\gamma}{4}}), \\ (\widetilde{B}G)_{\hat{i}\hat{i}} &= \frac{(z-\omega_B)\omega_B}{|\xi_i|^2 - \omega_B^2} + O_{\prec}(N^{-\frac{\gamma}{4}}). \end{aligned} \tag{5.95}$$

This together with (5.34) leads to the following estimates for $j = i, \hat{i}$:

$$\begin{aligned} -\tau_1(G)(\widetilde{B}G)_{ij} + \tau_1(G\widetilde{B})G_{ij} &= O_{\prec}(N^{-\frac{\gamma}{4}}), \\ -\tau_2(G)(\widetilde{B}G)_{\hat{i}j} + \tau_2(G\widetilde{B})G_{\hat{i}j} &= O_{\prec}(N^{-\frac{\gamma}{4}}), \end{aligned}$$

which together with (5.93) implies

$$\begin{aligned} \big(1 - \tau_1(G\widetilde{B})\big)T_{ij} + \tau_1(G)\tilde{\sigma}_i T_{\hat{i}j} &= O_{\prec}(N^{-\frac{\gamma}{4}}), \\ \big(1 - \tau_2(G\widetilde{B})\big)T_{\hat{i}j} + \tau_2(G)\tilde{\sigma}_i^* T_{ij} &= O_{\prec}(N^{-\frac{\gamma}{4}}), \qquad j = i, \hat{i}. \end{aligned} \tag{5.96}$$

Solving T_{ij} from the equations in (5.96), we get

$$\big(\big(1 - \tau_1(G\widetilde{B})\big)\big(1 - \tau_2(G\widetilde{B})\big) - |\sigma_i|^2\tau_1(G)\tau_2(G)\big)T_{ij} = O_{\prec}(N^{-\frac{\gamma}{4}}). \tag{5.97}$$

Using the assumption on Λ_T in (5.27), and also (5.34), we obtain from (5.97) that

$$\big(\big(1 + (\omega_B - z)m_{\mu_A \boxplus \mu_B}\big)^2 - |\sigma_i|^2 m_{\mu_A \boxplus \mu_B}^2\big)T_{ij} = O_{\prec}(N^{-\frac{\gamma}{4}}). \tag{5.98}$$

Further, observe that

$$\begin{aligned}(1+(\omega_B-z)m_{\mu_A\boxplus\mu_B})^2-|\sigma_i|^2m_{\mu_A\boxplus\mu_B}^2\\ =m_{\mu_A\boxplus\mu_B}^2(\omega_A-|\sigma_i|)(\omega_A+|\sigma_i|),\end{aligned}\tag{5.99}$$

which follows from the second equation in (2.5) with $(\mu_1,\mu_2)=(\mu_A,\mu_B)$. Then by (A.4) and the fact $m_{\mu_A\boxplus\mu_B}=m_{\mu_A}(\omega_B)$, we see that $|T_{ij}|\prec N^{-\frac{\gamma}{4}}$ for $j=i,\hat{i}$. Analogously, one can show $|T_{\hat{i}j}|\prec N^{-\frac{\gamma}{4}}$. This completes the proof of the crude bound (5.92).

With (5.92), we can now proceed to the proof of (5.29). We consider the average of $\mathcal{P}_{ii}$ over $i\in[\![1,N]\!]$, and use (5.28) to obtain

$$\Upsilon_1\cdot\frac{1}{N}\sum_{i=1}^N(G_{ii}+T_{ii})=\frac{1}{N}\sum_{i=1}^N\mathcal{P}_{ii}=O_\prec(\Psi).\tag{5.100}$$

By the first estimate in (5.34), the fact $m_{\mu_A\boxplus\mu_B}=m_{\mu_A}(\omega_B)$, the lower bound on $\operatorname{Im}\omega_B$ in (A.4), and also the crude bound (5.92), we can see that

$$\left|\frac{1}{\frac{1}{N}\sum_{i=1}^N(G_{ii}+T_{ii})}\right|=\left|\frac{1}{m_{\mu_A}(\omega_B)+O_\prec(N^{-\frac{\gamma}{4}})}\right|\prec 1.\tag{5.101}$$

Then the first estimate in (5.29) follows from (5.100) and (5.101) immediately. The second one can be verified similarly.

Finally, using (5.28) and (5.29), we can prove (5.30) as follows. Recall the definition in (5.19). Applying (5.27)–(5.29), we obtain, for $j=i,\hat{i}$,

$$(\widetilde{B}G)_{ij}=G_{ij}\frac{\tau_1(\widetilde{B}G)}{\tau_1(G)}+O_\prec(\Psi),\qquad(\widetilde{B}G)_{\hat{i}j}=G_{\hat{i}j}\frac{\tau_2(\widetilde{B}G)}{\tau_2(G)}+O_\prec(\Psi).\tag{5.102}$$

Using (5.94) and (5.102), we get the following system of equations:

$$\begin{aligned}1-\xi_iG_{\hat{i}i}+\omega_{B,1}^cG_{ii}=O_\prec(\Psi),\qquad&-\xi_iG_{\hat{i}\hat{i}}+\omega_{B,1}^cG_{i\hat{i}}=O_\prec(\Psi),\\ -\bar{\xi}_iG_{ii}+\omega_{B,2}^cG_{\hat{i}i}=O_\prec(\Psi),\qquad&1-\bar{\xi}_iG_{i\hat{i}}+\omega_{B,2}^cG_{\hat{i}\hat{i}}=O_\prec(\Psi),\end{aligned}\tag{5.103}$$

where we used the notation introduced in (8.20). Solving (5.103), we find

$$\begin{aligned}G_{ii}&=\frac{\omega_{B,2}^c}{|\xi_i|^2-\omega_{B,1}^c\omega_{B,2}^c}+O_\prec(\Psi),\\ G_{i\hat{i}}&=\frac{\xi_i}{|\xi_i|^2-\omega_{B,1}^c\omega_{B,2}^c}+O_\prec(\Psi),\\ G_{\hat{i}i}&=\frac{\bar{\xi}_i}{|\xi_i|^2-\omega_{B,1}^c\omega_{B,2}^c}+O_\prec(\Psi),\\ G_{\hat{i}\hat{i}}&=\frac{\omega_{B,1}^c}{|\xi_i|^2-\omega_{B,1}^c\omega_{B,2}^c}+O_\prec(\Psi).\end{aligned}\tag{5.104}$$

From (5.34), we see that

$$\omega_{B,a}^c = \omega_B + O_\prec(N^{-\frac{\gamma}{4}}), \qquad a = 1, 2. \tag{5.105}$$

The first estimate of (5.30) could be verified from (5.104), if we could show

$$\omega_{B,a}^c = \omega_B^c + O_\prec(\Psi), \qquad a = 1, 2. \tag{5.106}$$

To this end, we use $\tau_1(G(z)) = \tau_2(G(z))$; cf. (4.27). From (8.20) and (4.27), we also have

$$\omega_{B,1}^c + \omega_{B,2}^c = 2\omega_B^c. \tag{5.107}$$

Then, averaging the first and the fourth equations of (5.104) over $i \in [\![1, N]\!]$, we get

$$\omega_{B,2}^c \frac{1}{N}\sum_{i=1}^N \frac{1}{|\xi_i|^2 - \omega_{B,1}^c\omega_{B,2}^c} = \omega_{B,1}^c \frac{1}{N}\sum_{i=1}^N \frac{1}{|\xi_i|^2 - \omega_{B,1}^c\omega_{B,2}^c} + O_\prec(\Psi), \tag{5.108}$$

where we also used (4.27). We further claim that

$$\left(\frac{1}{N}\sum_{i=1}^N \frac{1}{|\xi_i|^2 - \omega_{B,1}^c\omega_{B,2}^c}\right)^{-1} \prec 1, \tag{5.109}$$

which together with (5.108) implies that

$$\omega_{B,2}^c = \omega_{B,1}^c + O_\prec(\Psi). \tag{5.110}$$

Combining (5.110) with (5.107), we get (5.106). Hence, it suffices to show (5.109). To this end, we use (5.105). Then we have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \frac{1}{|\xi_i|^2 - \omega_{B,1}^c\omega_{B,2}^c} &= \frac{1}{N}\sum_{i=1}^N \frac{1}{|\xi_i|^2 - \omega_B^2 + O_\prec(N^{-\frac{\gamma}{4}})} \\ &= \frac{1}{N}\sum_{i=1}^N \frac{1}{|\xi_i|^2 - \omega_B^2} + O_\prec(N^{-\frac{\gamma}{4}}) \\ &= \omega_B^{-1} m_{\mu_A}(\omega_B) + O_\prec(N^{-\frac{\gamma}{4}}),\end{aligned}$$

where in the first step above, we used the upper bound of $|\omega_B|$ in (A.4); in the second step, we used again the fact that $|\xi_i|^2 - \omega_B^2$ is away from 0 due to the lower bound of $\operatorname{Im}\omega_B$ in (A.4); and the last step follows from (5.33). Then the fact $\|A\| \le C$ (cf. (4.4)), the lower bound of $\operatorname{Im}\omega_B$ and the upper bound on $|\omega_B|$ in (A.4), we can get (5.109). Hence, we conclude the proof of the first estimate of (5.30).

For the second estimate in (5.30), we need to go through the proof of (5.92) again, but this time with the a priori input (5.27) replaced by the first estimate of (5.30). Therefore, with (5.30), we can get

$$\big((1 + (\omega_B^c - z)m_A(\omega_B^c))^2 - |\sigma_i|^2(m_A(\omega_B^c))^2\big)T_{ij} = O_\prec(\Psi), \tag{5.111}$$

which is the analogue of (5.98). Then, by the estimates in (5.34) and the definition in (5.22), it is not difficult to check that the coefficient of T_{ij} above can be approximated by (5.99), up to an error $O_{\prec}(N^{-\frac{\gamma}{4}})$. Hence, we can improve the estimate to $|T_{ij}| \prec \Psi$ for $j = i, \hat{i}$. Similarly, we can prove the same bound for $T_{\hat{i}j}$. This completes the second estimate of (5.30). Hence, we conclude the proof of Theorem 5.2. □

5.5. *Continuity argument*: *Proof of Theorem* 5.1. Having derived Theorem 5.2, we prove Theorem 5.1 using a continuity argument similar to [23].

PROOF OF THEOREM 5.1. First we show that $\Lambda_{\mathrm{d}}^{c}(z)$ in (5.30) can be replaced by $\Lambda_{\mathrm{d}}(z)$. This means, we have to control the difference between (ω_A, ω_B) and (ω_A^c, ω_B^c) as described in (5.26); this estimate will follow from the stability of the system $\Phi_{\mu_A,\mu_B}(\omega_A, \omega_B, z) = 0$ (cf. (4.29) with $(\mu_1, \mu_2) = (\mu_A, \mu_B)$). We will use the dual pair of subordination equations, that is, when we analyze $\mathcal{H}$ instead of H. Recall the notation introduced in (4.23), and also $\widetilde{\Lambda}_{\mathrm{d}}$ and $\widetilde{\Lambda}_T$ as the analogue of Λ_{d} and Λ_T, respectively, see the explanation around (5.23). For any $\delta \in [0, 1]$ and $z \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$, we introduce the following event:

$$\Theta(z, \delta) := \big\{\Lambda_{\mathrm{d}}(z) \le \delta, \widetilde{\Lambda}_{\mathrm{d}}(z) \le \delta, \Lambda_T(z) \le 1, \widetilde{\Lambda}_T(z) \le 1\big\}. \tag{5.112}$$

With the above notation, we have the following lemma.

LEMMA 5.5. *Suppose that the assumptions in Theorem* 4.3 *hold. Let* $\eta_{\mathrm{M}} > 0$ *be a sufficiently large constant and* $\gamma > 0$ *be a small constant in the definition* (5.1). *For any* ε *with* $0 < \varepsilon \le \frac{\gamma}{8}$ *and for any* $D > 0$, *there exists a positive integer* $N_2(D, \varepsilon)$ *such that the following holds*: *For any fixed* $z \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$ *there exists an event* $\Omega(z) \equiv \Omega(z, D, \varepsilon)$ *with*

$$\mathbb{P}\big(\Omega(z)\big) \ge 1 - N^{-D} \qquad \forall N \ge N_2(D, \varepsilon), \tag{5.113}$$

such that if the estimate

$$\mathbb{P}\big(\Theta\big(z, N^{-\frac{\gamma}{4}}\big)\big) \ge 1 - N^{-D}\big(1 + N^5(\eta_{\mathrm{M}} - \eta)\big), \qquad \eta = \operatorname{Im} z, \tag{5.114}$$

holds for all $D > 0$ *and* $N \ge N_1(D, \gamma, \varepsilon)$, *for some threshold* $N_1(D, \gamma, \varepsilon)$, *then we also have*

$$\Theta\big(z, N^{-\frac{\gamma}{4}}\big) \cap \Omega(z) \subset \Theta\bigg(z, \frac{N^{\varepsilon}}{\sqrt{N\eta}}\bigg), \tag{5.115}$$

for all $N \ge N_3(D, \gamma, \varepsilon) := \max\{N_1(D, \gamma, \varepsilon), N_2(D, \varepsilon)\}$.

PROOF. In this proof we fix $z \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$. According to the definition of $\prec$ in Definition 1.6, we see from the assumption (5.114) that

$$\Lambda_{\mathrm{d}}(z) \prec N^{-\frac{\gamma}{4}}, \qquad \widetilde{\Lambda}_{\mathrm{d}}(z) \prec N^{-\frac{\gamma}{4}}, \qquad \Lambda_T(z) \prec 1, \qquad \widetilde{\Lambda}_T(z) \prec 1. \tag{5.116}$$

We apply Theorem 5.2; by the estimates on Λ_{d}^c and on Λ_T in (5.30) and their analogues for $\widetilde{\Lambda}_{\mathrm{d}}^c$ and $\widetilde{\Lambda}_T$, we have

$$\Lambda_{\mathrm{d}}^c(z) \prec \Psi, \qquad \widetilde{\Lambda}_{\mathrm{d}}^c(z) \prec \Psi, \qquad \Lambda_T(z) \prec \Psi, \qquad \widetilde{\Lambda}_T(z) \prec \Psi. \tag{5.117}$$

Now we state the conclusions in (5.117) in a more explicit quantitative form, with the quantitative assumption (5.114). To this end, we need a more quantitative version of Lemma 5.3. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a smooth cutoff function s.t.

$$\begin{aligned} &\varphi(x) = 1 \qquad \text{if } |x| \le K, \qquad \varphi(x) = 0 \qquad \text{if } |x| \ge 2K, \\ &\sup_{x\in\mathbb{R}} |\varphi'(x)| \le CK^{-1} \end{aligned} \tag{5.118}$$

for some sufficiently large constant $K > 0$. Let

$$\begin{aligned} \Gamma_i \equiv \Gamma_i(z) &:= \sum_{a,b=i,\hat{i}} \big(|G_{ab}|^2 + |\mathcal{G}_{ab}|^2 + |T_{ab}|^2 + |\mathcal{T}_{ab}|^2\big) \\ &\quad + \sum_{a=1,2} \big(|\tau_a(G)|^2 + |\tau_a(\widetilde{B}G)|^2 + |\tau_a(G\widetilde{B})|^2 + |\tau_a(\widetilde{B}G\widetilde{B})|^2\big). \end{aligned} \tag{5.119}$$

Note that for a given i, all the a priori bounds we needed in the proof of Lemma 5.3 are the $O_\prec(1)$ bound for G_{ab}, $\mathcal{G}_{ab}$, T_{ab}, $\mathcal{T}_{ab}$ with $a, b = i, \hat{i}$ and the tracial quantities in (5.119). The $O_\prec(1)$ bound for $(XGY)_{ab}$ with $X, Y = \hat{I}$ or $\widetilde{B}$ were also used (see $(\widetilde{B}X\widetilde{B})_{ii}$ in (5.90) for instance), but they can be derived from the bound of G_{ab}'s by using (4.16). Recall the definitions of $\mathfrak{m}_i$ and $\mathfrak{n}_i$ in (5.37). We now introduce modifications of $\mathfrak{m}_i$ and $\mathfrak{n}_i$ by setting

$$\widetilde{\mathfrak{m}}_i(p,q) := \mathfrak{m}_i(p,q)(\varphi(\Gamma_i))^{p+q}, \qquad \widetilde{\mathfrak{n}}_i(p,q) := \mathfrak{n}_i(p,q)(\varphi(\Gamma_i))^{p+q}.$$

In addition, for any $\varepsilon' > 0$, let $\widehat{\Omega}(z) = \widehat{\Omega}(z, \varepsilon')$ be the event that all the concentration estimates of the components or quadratic forms of $\boldsymbol{h}_i^u$ and $\boldsymbol{h}_i^v$ in the proof of Lemma 5.3 hold with precision $N^{\varepsilon'}$. For instance, we used the large deviation bound (A.1) to bound $(\boldsymbol{k}_i^u)^* \widetilde{B}^{\langle i\rangle} \boldsymbol{k}_i^v$ in (5.41) by $O_\prec(N^{-\frac{1}{2}})$, in the proof of Lemma 5.3. Now we can bound it more quantitatively by $\frac{N^{\varepsilon'}}{\sqrt{N}}$ on $\widehat{\Omega}(z)$. Now we claim that

$$\begin{aligned} \mathbb{E}[\widetilde{\mathfrak{m}}_i(p,p)] &= \mathbb{E}\big[\mathfrak{c}_1 \widetilde{\mathfrak{m}}_i(p-1,p)\big] + \mathbb{E}\big[\mathfrak{c}_2 \widetilde{\mathfrak{m}}_i(p-2,p)\big] \\ &\quad + \mathbb{E}\big[\mathfrak{c}_3 \widetilde{\mathfrak{m}}_i(p-1,p-1)\big] \end{aligned} \tag{5.120}$$

with some random variables $\mathfrak{c}_1$, $\mathfrak{c}_2$, $\mathfrak{c}_3$, satisfying

$$|\mathfrak{c}_1| \le C\frac{N^{\varepsilon'}}{\sqrt{N\eta}}, \qquad |\mathfrak{c}_2| \le C\frac{N^{2\varepsilon'}}{N\eta}, \qquad |\mathfrak{c}_3| \le C\frac{N^{2\varepsilon'}}{N\eta} \qquad \text{on } \widehat{\Omega}(z), \tag{5.121}$$

for some positive constant C which may depend on K in (5.118). In addition, the $\mathfrak{c}_i$'s also admit trivial deterministic bounds of order η^{-k}, for some constant $k > 0$. Moreover, for any $D' > 0$, there exists $N(D', \varepsilon')$, such that if $N \geq N(D', \varepsilon')$

$$\mathbb{P}\big(\widehat{\Omega}(z)\big) \geq 1 - N^{-D'}.$$

Observe that (5.120) is just a more explicit version of (5.39), considering that $\widehat{\Omega}(z)$ holds with high probability. The proof of the more quantitative estimate (5.120) with (5.121) is basically the same as the proof of the nonquantitative one in (5.39).

The price for introducing $\varphi(\Gamma_i)$ into $\widetilde{\mathfrak{m}}_i$ is that it creates additional terms in the integration by parts. However, they are absorbed into the first term in the right side of (5.120). For instance, in the analogue of the step (5.56), except for replacing $\mathfrak{m}_i$ by $\widetilde{\mathfrak{m}}_i$, we will have an additional term

$$\frac{1}{N}\sum_k^{(i)} \mathbb{E}\bigg[\frac{1}{\|\boldsymbol{g}_i^u\|_2}\big(\hat{\boldsymbol{e}}_k^* \widetilde{B}^{\langle i\rangle} G \hat{\boldsymbol{e}}_i\big)\frac{\partial \varphi(\Gamma_i)}{\partial g_{ik}^u}\tau_1(G)\widetilde{\mathfrak{m}}_i(p-1,p)\bigg].$$

For example, one term of $\frac{\partial \varphi(\Gamma_i)}{\partial g_{ik}^u}$ is

$$\varphi'(\Gamma_i)\frac{\partial |G_{ii}|^2}{\partial g_{ik}^u} = \varphi'(\Gamma_i)\frac{\partial G_{ii}}{\partial g_{ik}^u}\overline{G}_{ii} + \varphi'(\Gamma_i)\frac{\partial \overline{G_{ii}}}{\partial g_{ik}^u}G_{ii}.$$

Using the second estimate in (5.75),

$$\frac{1}{N}\sum_k^{(i)} \hat{\boldsymbol{e}}_k^* \widetilde{B}^{\langle i\rangle} G \hat{\boldsymbol{e}}_i \frac{\partial |G_{ii}|^2}{\partial g_{ik}^u} = O\Big(\frac{N^{\varepsilon'}}{\sqrt{N\eta}}\Big) \qquad \text{on } \{\varphi'(\Gamma_i) \neq 0\} \cap \widehat{\Omega}(z).$$

It is also easy to check that the other terms in $\frac{\partial \varphi(\Gamma_i)}{\partial g_{ik}^u}$ give the same bound. Therefore, we have (5.120).

Using Young's inequality to (5.120), we can get

$$\begin{aligned}\mathbb{E}\big[\widetilde{\mathfrak{m}}_i(p,p)\big] &\leq C_p N^{2p\varepsilon'}\big(\mathbb{E}\big[|\mathfrak{c}_1|^{2p}\big] + \mathbb{E}\big[|\mathfrak{c}_2|^p\big] + \mathbb{E}\big[|\mathfrak{c}_3|^p\big]\big)\\ &\leq C_p N^{2p\varepsilon'}\bigg(\Big(\frac{N^{\varepsilon'}}{\sqrt{N\eta}}\Big)^{2p} + N^{-D'}\eta^{-2kp}\bigg),\end{aligned}$$

which implies by Markov's inequality that

$$\begin{aligned}&\mathbb{P}\bigg(|\mathcal{P}_{ii}\varphi(\Gamma_i)| \geq \frac{N^{\frac{\varepsilon}{4}}}{\sqrt{N\eta}}\bigg)\\ &\quad\leq C_p\bigg(\frac{N^{\frac{\varepsilon}{4}}}{\sqrt{N\eta}}\bigg)^{-2p} N^{2p\varepsilon'}\bigg(\Big(\frac{N^{\varepsilon'}}{\sqrt{N\eta}}\Big)^{2p} + N^{-D'}\eta^{-2kp}\bigg).\end{aligned} \tag{5.122}$$

For the given $\varepsilon > 0$ in Lemma 5.5, by first choosing $\varepsilon' = \varepsilon'(\varepsilon)$ to be smaller than $\frac{\varepsilon}{8}$, and then choosing $p = p(\varepsilon, D)$ to be sufficiently large, we get

$$C_p\bigg(\frac{N^{\frac{\varepsilon}{4}}}{\sqrt{N\eta}}\bigg)^{-2p} N^{2p\varepsilon'}\bigg(\frac{N^{\varepsilon'}}{\sqrt{N\eta}}\bigg)^{2p} \leq \frac{1}{2}N^{-D}. \tag{5.123}$$

Then, by further choosing $D' = D'(\varepsilon, D)$ sufficiently large, we can guarantee

$$C_p\Big(\frac{N^{\frac{\varepsilon}{4}}}{\sqrt{N\eta}}\Big)^{-2p} N^{2p\varepsilon'} N^{-D'} \eta^{-2kp} \le \frac{1}{2} N^{-D}. \tag{5.124}$$

With these choices of ε' and D', we now set $N_2(D, \varepsilon) := N(D', \varepsilon')$.

Further, by (5.122)–(5.124), there exists an event $\Omega(z)$, such that

$$\mathbb{P}(\Omega(z)) \ge 1 - N^{-D}, \qquad N \ge N_2(D, \varepsilon)$$

and

$$\big|\mathcal{P}_{ii}\varphi(\Gamma_i)\big| \le \frac{N^{\frac{\varepsilon}{4}}}{\sqrt{N\eta}} \qquad \text{on } \Omega(z).$$

This now implies that $|\mathcal{P}_{ii}| \le \frac{N^{\frac{\varepsilon}{4}}}{\sqrt{N\eta}}$ on $\Theta(z, N^{-\frac{\gamma}{4}}) \cap \Omega(z)$. Similarly, by working on $\widetilde{\mathfrak{n}}_i$, we can get $|\mathcal{K}_{ii}| \le \frac{N^{\frac{\varepsilon}{4}}}{\sqrt{N\eta}}$ on $\Theta(z, N^{-\frac{\gamma}{4}}) \cap \Omega(z)$.

The same bound can be obtained for $\mathcal{P}_{ij}$, $\mathcal{P}_{\hat{i}j}$, $\mathcal{K}_{ij}$ and $\mathcal{K}_{\hat{i}j}$ for $j = i, \hat{i}$. The remaining argument is the same as the proof of (5.30) in Theorem 5.2. The only change is, instead of the notation $\prec$, we use the deterministic $\le$, but restricting onto the event $\Theta(z, N^{-\frac{\gamma}{4}}) \cap \Omega(z)$.

More specifically, the quantitative proof of (5.117) yields that

$$\begin{aligned} \Lambda_{\mathrm{d}}^c(z) &\le \frac{N^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \qquad & \widetilde{\Lambda}_{\mathrm{d}}^c(z) &\le \frac{N^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \\ \Lambda_T(z) &\le \frac{N^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \qquad & \widetilde{\Lambda}_T(z) &\le \frac{N^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}} \end{aligned} \tag{5.125}$$

hold on the event $\Theta(z, N^{-\frac{\gamma}{4}}) \cap \Omega(z)$, for all $N \ge N_3(D, \gamma, \varepsilon)$.

Therefore, by the definitions of Λ_{d}^c and $\widetilde{\Lambda}_{\mathrm{d}}^c$, we have

$$\begin{aligned} \Big|G_{ii} - \frac{\omega_B^c}{|\xi_i|^2 - (\omega_B^c)^2}\Big| &\le \frac{N^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \\ \Big|\mathcal{G}_{ii} - \frac{\omega_A^c}{|\sigma_i|^2 - (\omega_A^c)^2}\Big| &\le \frac{N^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \\ \Big|G_{\hat{i}\hat{i}} - \frac{\omega_B^c}{|\xi_i|^2 - (\omega_B^c)^2}\Big| &\le \frac{N^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \\ \Big|\mathcal{G}_{\hat{i}\hat{i}} - \frac{\omega_A^c}{|\sigma_i|^2 - (\omega_A^c)^2}\Big| &\le \frac{N^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \end{aligned} \tag{5.126}$$

for all $i \in [\![1, N]\!]$, on the event $\Theta(z, N^{-\frac{\gamma}{4}}) \cap \Omega(z)$ for all $N \geq N_3(D, \gamma, \varepsilon)$. Averaging the above estimates over i, we obtain the system of equations

$$\begin{aligned} m_H(z) &= m_A\big(\omega_B^c(z)\big) + r_A(z), \\ m_H(z) &= m_B\big(\omega_A^c(z)\big) + r_B(z), \\ \omega_A^c(z) + \omega_B^c(z) &= z - \frac{1}{m_H(z)}, \end{aligned} \tag{5.127}$$

where the error terms $r_A(z)$ and $r_B(z)$ satisfy

$$|r_A(z)| \leq \frac{CN^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \qquad |r_B(z)| \leq \frac{CN^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}},$$

on the event $\Theta(z, N^{-\frac{\gamma}{4}}) \cap \Omega(z)$ for all $N \geq N_3(D, \gamma, \varepsilon)$. Here, the last equation in (5.127) follows from the definition (4.17) or (5.22). From the definition of $\Theta(z, \delta)$ in (5.112), (4.17) or (5.22), and the equations in (2.5) with $(\mu_1, \mu_2) = (\mu_A, \mu_B)$, it is not difficult to check that

$$|\omega_A^c - \omega_A| \leq CN^{-\frac{\gamma}{4}}, \qquad |\omega_B^c - \omega_B| \leq CN^{-\frac{\gamma}{4}}$$

hold on $\Theta(z, N^{-\frac{\gamma}{4}})$. In particular, with the help of (A.4), this guarantees that the imaginary parts of ω_A^c and ω_B^c are separated away from zero, hence so are $m_A(\omega_B^c)$ and $m_B(\omega_B^c)$. This allows us to rewrite (5.127) as

$$\big\|\Phi_{\mu_A,\mu_B}\big(\omega_A^c, \omega_B^c, z\big)\big\| = \widetilde{r}(z), \tag{5.128}$$

where $\widetilde{r}(z) = (\widetilde{r}_A(z), \widetilde{r}_B(z))'$ satisfy

$$|\widetilde{r}_A(z)| \leq \frac{CN^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \qquad |\widetilde{r}_B(z)| \leq \frac{CN^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}},$$

on the event $\Theta(z, N^{-\frac{\gamma}{4}}) \cap \Omega(z)$ for all $N \geq N_3(D, \gamma, \varepsilon)$. Applying the stability of the system $\Phi_{\mu_A,\mu_B}(\omega_A, \omega_B, z) = 0$ (see Theorem 4.1 of [2]), we obtain

$$|\omega_A^c - \omega_A| \leq \frac{CN^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \qquad |\omega_B^c - \omega_B| \leq \frac{CN^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \tag{5.129}$$

on the event $\Theta(z, N^{-\frac{\gamma}{4}}) \cap \Omega(z)$ for all $N \geq N_3(D, \gamma, \varepsilon)$. Substituting (5.129) into the definition of Λ_{d}^c and $\widetilde{\Lambda}_{\mathrm{d}}^c$, we see that the first two inequalities in (5.125) imply similar bounds for Λ_{d} and $\widetilde{\Lambda}_{\mathrm{d}}$. This completes the proof of Lemma 5.5. □

With Lemma 5.5, the remaining proof of Theorem 5.1 closely follows that for Theorem 2.5 in [3], so we will only sketch the argument. We start with the result with large $\eta = \eta_M$ for some large but fixed positive constant η_{M}. More specifically, from Lemma 8.1, we see that

$$\Lambda_{\mathrm{d}}(E + \mathrm{i}\eta_{\mathrm{M}}) \prec \frac{1}{\sqrt{N\eta_{\mathrm{M}}^4}}, \qquad \widetilde{\Lambda}_{\mathrm{d}}(E + \mathrm{i}\eta_{\mathrm{M}}) \prec \frac{1}{\sqrt{N\eta_{\mathrm{M}}^4}}, \tag{5.130}$$

for any fixed $E \in \mathbb{R}$. The second estimate in (5.130) can be obtained from Lemma 8.1 since one can apply this lemma to $\mathcal{H}$ as well. In addition, using the trivial bound $\|G\| \le \frac{1}{\eta}$ and inequality $|\boldsymbol{x}^* G \boldsymbol{y}| \le \|G\| \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2$, we also have

$$\Lambda_T(E + \mathrm{i}\eta_{\mathrm{M}}) \le \frac{1}{\eta_{\mathrm{M}}}, \qquad \widetilde{\Lambda}_T(E + \mathrm{i}\eta_{\mathrm{M}}) \le \frac{1}{\eta_{\mathrm{M}}}, \tag{5.131}$$

for any fixed $E \in \mathcal{B}_{\mu_\alpha \boxplus \mu_\beta}$. According to the definition of $\Theta(z, \delta)$ in (5.112), (5.130) and (5.131), we see that for any fixed $E \in \mathcal{B}_{\mu_\alpha \boxplus \mu_\beta}$ and $D > 0$,

$$\mathbb{P}\big(\Theta\big(E + \mathrm{i}\eta_{\mathrm{M}}, N^{-\frac{3\gamma}{8}}\big)\big) \ge 1 - N^{-D}, \tag{5.132}$$

holds for all $N \ge N_0(D, \gamma)$ for some positive integer $N_0(D, \gamma)$.

Starting with (5.132), we conduct a standard continuity argument, whose setup is best suited to our problem in the form presented in [3]. Specifically, we do a bootstrap by reducing η in very small steps, N^{-5} (say), starting from η_{M} and successively control the probability of the "good" events Θ. Recall the event $\Omega(z)$ in Lemma 5.5. The main task is to show for any fixed $E \in \mathcal{I}$ and any $\eta \in [\eta_{\mathrm{m}}, \eta_{\mathrm{M}}]$,

$$\Theta\big(E + \mathrm{i}\eta, N^{-\frac{3\gamma}{8}}\big) \cap \Omega\big(E + \mathrm{i}\big(\eta - N^{-5}\big)\big) \subset \Theta\big(E + \mathrm{i}\big(\eta - N^{-5}\big), N^{-\frac{3\gamma}{8}}\big), \tag{5.133}$$

which is the analogue of (7.20) of [3]. To see this inclusion, one first uses the Lipschitz continuity of the Green function, $\|G(z) - G(z')\| \le N^2 |z - z'|$, and of the subordination functions (cf. (A.4)) to obtain

$$\Theta\big(E + \mathrm{i}\eta, N^{-\frac{3\gamma}{8}}\big) \subset \Theta\big(E + \mathrm{i}\big(\eta - N^{-5}\big), N^{-\frac{\gamma}{4}}\big). \tag{5.134}$$

Then (5.134) together with (5.115) implies (5.133). Using (5.133) recursively, one goes from η_{M} down to η_{m}, step by step. The remaining proof of (5.25), based on (5.133) and Lemma 5.5, is the same as the counterpart in [3] (cf. (7.20)–(7.25) therein). We omit the details.

With (5.25), we can prove (5.26) in the sequel. The first two inequalities in (5.26) have already been proved in (5.129) with a fixed η, under (5.116). The uniformity then follows from (5.25) which holds uniformly on $\mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$. Then the last inequality in (5.26) follows from the first two, together with the last equation in (5.127) and the second equation in (2.5) with $(\mu_1, \mu_2) = (\mu_A, \mu_B)$. This completes the proof of Theorem 5.1. □

6. Strong law for small η. In this section we prove the strong law, that is, Theorem 4.3, for $z \in \mathcal{S}_{\mathcal{I}}(0, \eta_{\mathrm{M}})$. It suffices to work on the regime $z \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$ at first. The extension to $z \in \mathcal{S}_{\mathcal{I}}(0, \eta_{\mathrm{M}})$ will be easy. Our main task is to establish the fluctuation averaging for the quantities $\mathcal{P}_{ij}$ defined in (5.19).

LEMMA 6.1 (Fluctuation averaging). *Suppose that the assumptions in Theorem 4.3 hold. Let $\eta_{\mathrm{M}} > 0$ be any (large) constant and $\gamma > 0$ be any (small) constant*

in the definition of η_{m} (*cf.* (5.1)). *For any fixed integer* $p \geq 1$, *and deterministic numbers* $d_1, \ldots, d_N \in \mathbb{C}$ *satisfying* $\max_{i \in [\![1,N]\!]} |d_i| \leq 1$, *we have*

$$\begin{aligned} &\left|\frac{1}{N}\sum_{i=1}^{N} d_i \mathcal{P}_{ii}\right| \prec \Psi^2, \qquad \left|\frac{1}{N}\sum_{i=1}^{N} d_i \mathcal{P}_{\hat{i}i}\right| \prec \Psi^2, \\ &\left|\frac{1}{N}\sum_{i=1}^{N} d_i \mathcal{P}_{i\hat{i}}\right| \prec \Psi^2, \qquad \left|\frac{1}{N}\sum_{i=1}^{N} d_i \mathcal{P}_{\hat{i}\hat{i}}\right|^2 \prec \Psi^2 \end{aligned} \tag{6.1}$$

uniformly on $\mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$.

We will often use the following improvement of (5.34):

$$\begin{aligned} \tau_a(G) &= m_{\mu_A \boxplus \mu_B} + O_{\prec}(\Psi), \\ \tau_a(\widetilde{B}G) &= (z - \omega_B) m_{\mu_A \boxplus \mu_B} + O_{\prec}(\Psi), \\ \tau_a(G\widetilde{B}) &= (z - \omega_B) m_{\mu_A \boxplus \mu_B} + O_{\prec}(\Psi), \\ \tau_a(\widetilde{B}G\widetilde{B}) &= (\omega_B - z)\big(1 + (\omega_B - z) m_{\mu_A \boxplus \mu_B}\big) + O_{\prec}(\Psi), \qquad a = 1, 2, \end{aligned} \tag{6.2}$$

which can be proved in the same way as (5.34), but with the first inequality in (5.27) replaced by the first inequality in (5.25), as the input of the proof.

In the next Section 6.1, we will show how to prove Theorem 4.3 on $\mathcal{S}_{\mathcal{I}}(0, \eta_{\mathrm{M}})$ with the aid of Lemma 6.1. Then, in Section 6.2 we will prove Lemma 6.1.

6.1. *Proof of Theorem 4.3 on* $\mathcal{S}_{\mathcal{I}}(0, \eta_{\mathrm{M}})$. To prove the strong law from Lemma 6.1, first of all, we need to derive that the estimates

$$|\Upsilon_1| \prec \Psi^2, \qquad |\Upsilon_2| \prec \Psi^2 \tag{6.3}$$

hold uniformly on $\mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$. These are the strongest high probability bounds related to the Ward identities in (5.29). To see (6.3), we choose $d_i = 1$ for all $i \in [\![1, N]\!]$ in (6.1). From the definition of $\mathcal{P}_{ii}$ in (5.19), we get

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N} \mathcal{P}_{ii} = \frac{\Upsilon_1}{N}\sum_{i=1}^{N}(G_{ii} + T_{ii}) &= \Upsilon_1\left(\tau_1(G) + \frac{1}{N}\sum_{i=1}^{N} T_{ii}\right) \\ &= \Upsilon_1\big(m_{\mu_A \boxplus \mu_B} + O_{\prec}(\Psi)\big), \end{aligned} \tag{6.4}$$

where in the last step we used (6.2) and the third inequality in (5.25). Then, using the lower bound of $\operatorname{Im} m_{\mu_A \boxplus \mu_B} = \operatorname{Im} m_{\mu_A}(\omega_B)$ inherited from the lower bound of $\operatorname{Im} \omega_B$ in (A.4), and also the first bound in (6.1), we can easily see $|\Upsilon_1| \prec \Psi^2$ from (6.4). Similarly, we can also show $|\Upsilon_2| \prec \Psi^2$. Notice that *a posteriori* we could have defined $\mathcal{P}_{ij}$ in (5.19) without the last term involving Υ_a with $a = 1, 2$, since we are interested only up to $O_{\prec}(\Psi^2)$ precision. We do not, however, know how to prove directly that $\Upsilon_a = O_{\prec}(\Psi^2)$ without first proving a fluctuation averaging

result (6.1) involving the quantity $\mathcal{P}_{ij}$ with Υ_a. The correct choice of $\mathcal{P}_{ij}$ is the essential idea of the entire proof.

Plugging (6.3) back to the definition of $\mathcal{P}_{ii}, \mathcal{P}_{\hat{i}i}, \mathcal{P}_{i\hat{i}}$ and $\mathcal{P}_{\hat{i}\hat{i}}$ in (5.19), we obtain from (6.1)

$$\begin{aligned} &\left|\frac{1}{N}\sum_{i=1}^{N} d_i\big(G_{ij}\tau_1(\widetilde{B}G) - (\widetilde{B}G)_{ij}\tau_1(G)\big)\right| \prec \Psi^2, \\ &\left|\frac{1}{N}\sum_{i=1}^{N} d_i\big(G_{\hat{i}j}\tau_2(\widetilde{B}G) - (\widetilde{B}G)_{\hat{i}j}\tau_2(G)\big)\right| \prec \Psi^2, \qquad j = i, \hat{i}, \end{aligned} \tag{6.5}$$

for any deterministic numbers $d_1, \dots, d_N \in \mathbb{C}$ satisfying $|d_i| \lesssim 1$, which is a shorthand notation for $|d_i| \le C$ with some constant C. While Lemma 6.1 was formulated for $|d_i| \le 1$, it clearly holds as long as $|d_i| \lesssim 1$. Recall the notation introduced in (8.20). We claim that the following estimates can be derived from (5.94) and (6.5):

$$\begin{aligned} &\left|\frac{1}{N}\sum_{i=1}^{N} d_i\bigg(G_{ii} - \frac{\omega_{B,2}^c}{|\xi_i|^2 - \omega_{B,1}^c\omega_{B,2}^c}\bigg)\right| \prec \Psi^2, \\ &\left|\frac{1}{N}\sum_{i=1}^{N} d_i\bigg(G_{\hat{i}i} - \frac{\bar{\xi}_i}{|\xi_i|^2 - \omega_{B,1}^c\omega_{B,2}^c}\bigg)\right| \prec \Psi^2, \\ &\left|\frac{1}{N}\sum_{i=1}^{N} d_i\bigg(G_{\hat{i}\hat{i}} - \frac{\omega_{B,1}^c}{|\xi_i|^2 - \omega_{B,1}^c\omega_{B,2}^c}\bigg)\right| \prec \Psi^2, \\ &\left|\frac{1}{N}\sum_{i=1}^{N} d_i\bigg(G_{i\hat{i}} - \frac{\xi_i}{|\xi_i|^2 - \omega_{B,1}^c\omega_{B,2}^c}\bigg)\right| \prec \Psi^2. \end{aligned} \tag{6.6}$$

We derive the first estimate in (6.6), the others are proven similarly. We write

$$\begin{aligned} &\frac{1}{N}\sum_{i=1}^{N} d_i\bigg(G_{ii} - \frac{\omega_{B,2}^c}{|\xi_i|^2 - \omega_{B,1}^c\omega_{B,2}^c}\bigg) \\ &\quad = \frac{1}{N}\sum_{i=1}^{N} \frac{d_i}{|\xi_i|^2 - \omega_{B,1}^c\omega_{B,2}^c}\big(G_{ii}\big(|\xi_i|^2 - \omega_{B,1}^c\omega_{B,2}^c\big) - \omega_{B,2}^c\big). \end{aligned}$$

Applying Theorem 5.1 and (5.106) along its proof, it is easy to check that

$$\omega_{B,a}^c = \omega_B + O_\prec(\Psi), \qquad a = 1, 2, \tag{6.7}$$

hence

$$\omega_{B,1}^c\omega_{B,2}^c = \omega_B^2 + O_\prec(\Psi), \qquad G_{ii}\big(|\xi_i|^2 - \omega_{B,1}^c\omega_{B,2}^c\big) - \omega_{B,2}^c = O_\prec(\Psi). \tag{6.8}$$

Moreover, from the lower bound on $\operatorname{Im}\omega_B$ from (A.4) and the first estimate of (6.8), we have

$$\frac{1}{|\xi_i|^2-\omega_{B,1}^c\omega_{B,2}^c}=\frac{1}{|\xi_i|^2-\omega_B^2}+O_{\prec}(\Psi). \tag{6.9}$$

Then, in light of (A.4), (6.8) and (6.9), it suffices to check

$$\frac{1}{N}\sum_{i=1}^{N}d_i\big(G_{ii}\big(|\xi_i|^2-\omega_{B,1}^c\omega_{B,2}^c\big)-\omega_{B,2}^c\big)=O_{\prec}\big(\Psi^2\big), \tag{6.10}$$

for any deterministic numbers $d_1,\dots,d_N\in\mathbb{C}$ satisfying $|d_i|\lesssim 1$ (here we redefined d_i to $d_i/(|\xi_i|^2-\omega_B^2)$). Using (5.94), we can write

$$\begin{aligned}&G_{ii}\big(|\xi_i|^2-\omega_{B,1}^c\omega_{B,2}^c\big)-\omega_{B,2}^c\\&\quad=-\frac{\omega_{B,2}^c}{\tau_1(G)}\big((\widetilde{B}G)_{ii}\tau_1(G)-G_{ii}\tau_1(\widetilde{B}G)\big)\\&\qquad-\frac{\xi_i}{\tau_2(G)}\big((\widetilde{B}G)_{\hat{i}i}\tau_2(G)-G_{\hat{i}i}\tau_2(\widetilde{B}G)\big).\end{aligned} \tag{6.11}$$

Then, from (6.2) and (6.7), we see that

$$\begin{aligned}\frac{\omega_{B,2}^c}{\tau_1(G)}&=\frac{\omega_B}{m_{\mu_A\boxplus\mu_B}}+O_{\prec}(\Psi),\\\frac{\xi_i}{\tau_2(G)}&=\frac{\xi_i}{m_{\mu_A\boxplus\mu_B}}+O_{\prec}(\Psi),\\(\widetilde{B}G)_{ii}\tau_1(G)-G_{ii}\tau_1(\widetilde{B}G)&=O_{\prec}(\Psi),\\(\widetilde{B}G)_{\hat{i}i}\tau_2(G)-G_{\hat{i}i}\tau_2(\widetilde{B}G)&=O_{\prec}(\Psi),\end{aligned} \tag{6.12}$$

where the second line follows from (5.102). Thus combining (6.11), (6.12) and (6.5) yields (6.10), which implies (6.6) according to the discussion above.

Notice that in this argument it was essential that G_{ii} was approximated in (6.6) not by $\omega_B/(|\xi_i|^2-\omega_B^2)$ or by $\omega_B^c/(|\xi_i|^2-(\omega_B^c)^2)$ but by

$$G_{ii}\approx\frac{\omega_{B,2}^c}{|\xi_i|^2-\omega_{B,1}^c\omega_{B,2}^c},$$

since this latter approximation is precise up to $O_{\prec}(\Psi^2)$ after averaging, while the previous ones are *a priori* correct only with an error $O_{\prec}(\Psi)$.

Next we show that (6.6) nevertheless holds if we approximate G_{ii} by $\omega_B^c/(|\xi_i|^2-(\omega_B^c)^2)$. Choosing all $d_i=1$ in the first and third inequalities in (6.6) and applying (4.27), we note that

$$\omega_{B,a}^c=\omega_B^c+O_{\prec}\big(\Psi^2\big),\qquad a=1,2,$$

so the first approximation in (6.7) is actually one order better. Thus we get from (6.6) that

$$
\begin{aligned}
&\left|\frac{1}{N}\sum_{i=1}^{N} d_i\left(G_{ii}-\frac{\omega_B^c}{|\xi_i|^2-(\omega_B^c)^2}\right)\right| \prec \Psi^2,\\
&\left|\frac{1}{N}\sum_{i=1}^{N} d_i\left(G_{\hat{i}i}-\frac{\bar{\xi}_i}{|\xi_i|^2-(\omega_B^c)^2}\right)\right| \prec \Psi^2,\\
&\left|\frac{1}{N}\sum_{i=1}^{N} d_i\left(G_{\hat{i}\hat{i}}-\frac{\omega_B^c}{|\xi_i|^2-(\omega_B^c)^2}\right)\right| \prec \Psi^2,\\
&\left|\frac{1}{N}\sum_{i=1}^{N} d_i\left(G_{i\hat{i}}-\frac{\xi_i}{|\xi_i|^2-(\omega_B^c)^2}\right)\right| \prec \Psi^2.
\end{aligned}
\tag{6.13}
$$

Further, recalling the definitions of $\mathcal{H}$ and $\mathcal{G}$ in (4.23). Switching the roles of A and B, and also the roles of U and U^* in the above discussions, we have

$$
\begin{aligned}
&\left|\frac{1}{N}\sum_{i=1}^{N} d_i\left(\mathcal{G}_{ii}-\frac{\omega_A^c}{|\sigma_i|^2-(\omega_A^c)^2}\right)\right| \prec \Psi^2,\\
&\left|\frac{1}{N}\sum_{i=1}^{N} d_i\left(\mathcal{G}_{\hat{i}i}-\frac{\bar{\sigma}_i}{|\sigma_i|^2-(\omega_A^c)^2}\right)\right| \prec \Psi^2,\\
&\left|\frac{1}{N}\sum_{i=1}^{N} d_i\left(\mathcal{G}_{\hat{i}\hat{i}}-\frac{\omega_A^c}{|\xi_i|^2-(\omega_A^c)^2}\right)\right| \prec \Psi^2,\\
&\left|\frac{1}{N}\sum_{i=1}^{N} d_i\left(\mathcal{G}_{i\hat{i}}-\frac{\sigma_i}{|\sigma_i|^2-(\omega_A^c)^2}\right)\right| \prec \Psi^2.
\end{aligned}
\tag{6.14}
$$

Applying (6.13) and (6.14) to average over the diagonal entries of the Green functions G and $\mathcal{G}$, and also using the fact $\operatorname{tr} G(z)=\operatorname{tr}\mathcal{G}(z)=m_H(z)$, we see that

$$
\begin{aligned}
m_H(z) &= \int_{\mathbb{R}} \frac{\omega_B^c}{x^2-(\omega_B^c)^2}\,\mathrm{d}\mu_{\Xi}(x)+O_{\prec}(\Psi^2)\\
&= \int_{\mathbb{R}} \frac{\omega_A^c}{x^2-(\omega_A^c)^2}\,\mathrm{d}\mu_{\Sigma}(x)+O_{\prec}(\Psi^2).
\end{aligned}
$$

From this, using

$$
\frac{\omega_B^c}{x^2-(\omega_B^c)^2}=\frac{1}{2}\left[\frac{1}{x-\omega_B^c}+\frac{1}{-x-\omega_B^c}\right],
$$

we can get

$$
m_H(z)=m_A\big(\omega_B^c(z)\big)+O_{\prec}\big(\Psi^2\big)=m_B\big(\omega_A^c(z)\big)+O_{\prec}\big(\Psi^2\big),
\tag{6.15}
$$

where we used the fact $\mu_A \equiv \mu_\Xi^{\mathrm{sym}}$ and $\mu_B \equiv \mu_\Sigma^{\mathrm{sym}}$, in light of (4.14). In addition, we also have (4.18). Summarizing these estimates, we have $\Phi_{\mu_A,\mu_B}(\omega_A^c, \omega_B^c, z) = O_\prec(\Psi^2)$, that is, compared with (5.128), we improved the error in the approximate subordination equations.

Similar to the proof of Lemma 5.5, we use the stability of the system $\Phi_{\mu_A,\mu_B}(\omega_A, \omega_B, z) = 0$ again, but with the improved error Ψ^2. We also note that the estimates from Theorem 5.1 and Lemma 6.1 used in the above discussion hold uniformly on $\mathcal{S}_\mathcal{I}(\eta_\mathrm{m}, \eta_\mathrm{M})$. Hence, we can conclude the proof of Theorem 4.3 on $\mathcal{S}_\mathcal{I}(\eta_\mathrm{m}, \eta_\mathrm{M})$.

At the end, we extend (4.11) from $\mathcal{S}_\mathcal{I}(\eta_\mathrm{m}, \eta_\mathrm{M})$ to $\mathcal{S}_\mathcal{I}(0, \eta_\mathrm{M})$. The extension relies on a standard use of the monotonicity of the Green function: For all $i \in [\![1, N]\!]$ and $j = i$ or $\hat{i}$, we have

$$|G'_{jj}(z)| = \left|\sum_{k=1}^{2N} G_{jk}(z)G_{kj}(z)\right| \le \sum_{k=1}^{2N} |G_{jk}(z)|^2 = \frac{\operatorname{Im} G_{jj}(z)}{\eta},$$

where the last step follows from the spectral decomposition. In addition, note that the function $s \mapsto s \operatorname{Im} G_{jj}(E + \mathrm{i}s)$ is monotonically increasing. This implies that for any $\eta \in (0, \eta_\mathrm{m}]$,

$$\begin{aligned} |G_{jj}(E + \mathrm{i}\eta) - G_{jj}(E + \mathrm{i}\eta_\mathrm{m})| &\le \int_\eta^{\eta_\mathrm{m}} \frac{s \operatorname{Im} G_{jj}(E + \mathrm{i}s)}{s^2}\,\mathrm{d}s \\ &\le 2\frac{\eta_\mathrm{m}}{\eta} \operatorname{Im} G_{jj}(E + \mathrm{i}\eta_\mathrm{m}) \\ &\le C\frac{N^\gamma}{N\eta} \le CN^\gamma\Psi^2, \end{aligned} \tag{6.16}$$

with high probability, for any $E \in \mathcal{I}$. Here, we used $|G_{jj}(E + \mathrm{i}\eta_\mathrm{m})| \prec 1$ which follows from the first bound in (5.25). On the other hand, for any $i \in [\![1, N]\!]$, we also have

$$\left|\frac{\omega_B(E + \mathrm{i}\eta)}{|\xi_i|^2 - \omega_B^2(E + \mathrm{i}\eta)} - \frac{\omega_B(E + \mathrm{i}\eta_\mathrm{m})}{|\xi_i|^2 - \omega_B^2(E + \mathrm{i}\eta_\mathrm{m})}\right| \le C(\eta_\mathrm{m} - \eta) \le \Psi^2, \tag{6.17}$$

$\eta \in (0, \eta_\mathrm{m}]$, $E \in \mathcal{I}$, for sufficiently small γ, which follows from the upper bound of $\omega'_B(z)$, the lower bound of $|\xi_i|^2 - \omega_B^2(z)$ which follows from the lower bound of $\operatorname{Im}\omega_B$, and also the upper bound of ω_B, in Lemma A.2. Combining (6.16) and (6.17), and using (5.33), we conclude that (4.11) holds uniformly on $\mathcal{S}_\mathcal{I}(0, \eta_\mathrm{M})$. This completes the proof of Theorem 4.3 on $\mathcal{S}_\mathcal{I}(0, \eta_\mathrm{M})$.

Hence, what remains is to prove Lemma 6.1.

6.2. *Proof of Lemma 6.1.* Since the proofs for the four estimates in (6.1) are nearly the same, we only present the details for the first one. First of all, from (5.25) and (5.29) we have

$$|T_{ii}\Upsilon_1| \prec \Psi^2. \tag{6.18}$$

Hence, it suffices to bound the weighted average of the following slight modifications of $\mathcal{P}_{ii}$'s:

$$\mathcal{Q}_{ii} \equiv \mathcal{Q}_{ii}(z) := (\widetilde{B}G)_{ii}\tau_1(G) - G_{ii}\tau_1(\widetilde{B}G) + G_{ii}\Upsilon_1, \qquad i \in [\![1, N]\!]. \tag{6.19}$$

Then we introduce the notation

$$\mathfrak{m}(k, l) := \left(\frac{1}{N}\sum_{i=1}^{N} d_i \mathcal{Q}_{ii}\right)^k \left(\frac{1}{N}\sum_{i=1}^{N} \overline{d_i \mathcal{Q}_{ii}}\right)^l.$$

Similar to Lemma 5.3, the main technical task is the following recursive moment estimate.

THEOREM 6.2 (Recursive moment estimate). *Suppose that the assumptions in Theorem 4.3 hold. Let $\eta_{\mathrm{M}} > 0$ be any (large) constant and $\gamma > 0$ in (5.1) be any (small) constant. For any fixed integer $p \geq 1$, we have*

$$\begin{aligned}\mathbb{E}\big[\mathfrak{m}(p, p)\big] = \mathbb{E}\big[O_{\prec}(\Psi^2)\mathfrak{m}(p-1, p)\big] + \mathbb{E}\big[O_{\prec}(\Psi^4)\mathfrak{m}(p-2, p)\big] \\ + \mathbb{E}\big[O_{\prec}(\Psi^4)\mathfrak{m}(p-1, p-1)\big],\end{aligned} \tag{6.20}$$

uniformly on $\mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$, where we made the convention $\mathfrak{m}(0, 0) = 1$ and $\mathfrak{m}(-1, 1) = 0$ if $p = 1$.

The reason why we prefer to work with $\mathcal{Q}_{ii}$ instead of $\mathcal{P}_{ii} = \mathcal{Q}_{ii} + T_{ii}\Upsilon_i$ is as follows. To prove Theorem 6.2, we will follow a similar strategy as the proof of Lemma 5.3. In Lemma 5.3 and its proof, we worked on $\mathcal{P}_{ii}$ directly. The derivative $\frac{\partial T_{ii}}{\partial g^u_{ik}}$ was necessary for the proof of Lemma 5.3; cf. (5.75). However, in the proof of Theorem 6.2, we would need to consider the derivative $\frac{\partial T_{ii}}{\partial g_{jk}}$ for all $j \neq i$ if we carry the term $T_{ii}\Upsilon_1$ from $\mathcal{P}_{ii}$ in the discussion. Unfortunately, the dependence of the factor $(\boldsymbol{k}^u_i)^*$ in T_{ii} (cf. (5.16)) on g^u_{jk} for $j \neq i$ is difficult to capture. On the other hand, at this stage of the proof we already have the bound (6.18) available and this allows us to drop the term $T_{ii}\Upsilon_1$ from the beginning.

With the aid of Theorem 6.2, one can prove Lemma 6.1.

PROOF OF LEMMA 6.1. Similar to the proof of (5.28) for $\mathcal{P}_{ii}$ from Lemma 5.3, one can apply Young's inequality to (6.20) and get $|\frac{1}{N}\sum_{i=1}^{N}\mathcal{Q}_{ii}| \prec \Psi^2$, which together with (6.18) implies the first bound in (6.1). The other three in (6.1) can be verified analogously. Hence, we completed the proof of Lemma 6.1. □

PROOF OF THEOREM 6.2. Hence, we start with the averaged analogue of (5.71), but with $\mathcal{P}_{ii}$'s replaced by $\mathcal{Q}_{ii}$'s. In particular, the term T_{ii} is missing.

Following the proof of (5.71) with these modifications, we obtain

$$
\begin{aligned}
&\mathbb{E}\big[\mathfrak{m}(p,p)\big]\\
&\quad=\frac{1}{N}\sum_i d_i\mathbb{E}\bigg[\frac{1}{\|\boldsymbol{g}_i^u\|_2}\bigg(\mathring{T}_{ii}-\frac{1}{N}\sum_k^{(i)}\frac{\partial(\hat{\boldsymbol{e}}_k^*G\hat{\boldsymbol{e}}_i)}{\partial g_{ik}^u}\bigg)\\
&\qquad\times\tau_1(\widetilde{B}G)\mathfrak{m}(p-1,p)\bigg]\\
&\qquad-\frac{1}{N^2}\sum_i\sum_k^{(i)}d_i\mathbb{E}\bigg[\frac{\partial\|\boldsymbol{g}_i^u\|_2^{-1}}{\partial g_{ik}^u}\hat{\boldsymbol{e}}_k^*\widetilde{B}^{\langle i\rangle}G\hat{\boldsymbol{e}}_i\tau_1(G)\mathfrak{m}(p-1,p)\bigg]\\
&\qquad-\frac{1}{N^2}\sum_i\sum_k^{(i)}d_i\mathbb{E}\bigg[\frac{\partial\tau_1(G)}{\partial g_{ik}^u}\frac{1}{\|\boldsymbol{g}_i^u\|_2}\hat{\boldsymbol{e}}_k^*\widetilde{B}^{\langle i\rangle}G\hat{\boldsymbol{e}}_i\mathfrak{m}(p-1,p)\bigg]\\
&\qquad-\frac{p-1}{N^2}\sum_i\sum_k^{(i)}d_i\mathbb{E}\bigg[\frac{1}{\|\boldsymbol{g}_i^u\|_2}\hat{\boldsymbol{e}}_k^*\widetilde{B}^{\langle i\rangle}G\hat{\boldsymbol{e}}_i\tau_1(G)\\
&\qquad\times\bigg(\frac{1}{N}\sum_j d_j\frac{\partial\mathcal{Q}_{jj}}{\partial g_{ik}^u}\bigg)\mathfrak{m}(p-2,p)\bigg]\\
&\qquad-\frac{p}{N^2}\sum_i\sum_k^{(i)}d_i\mathbb{E}\bigg[\frac{1}{\|\boldsymbol{g}_i^u\|_2}\hat{\boldsymbol{e}}_k^*\widetilde{B}^{\langle i\rangle}G\hat{\boldsymbol{e}}_i\tau_1(G)\\
&\qquad\times\bigg(\frac{1}{N}\sum_j \bar{d}_j\frac{\partial\overline{\mathcal{Q}_{jj}}}{\partial g_{ik}^u}\bigg)\mathfrak{m}(p-1,p-1)\bigg]\\
&\qquad+\frac{1}{N}\sum_i d_i\mathbb{E}\bigg[\bigg(\varepsilon_{i1}\tau_1(G)-\frac{\varepsilon_{i4}+\varepsilon_{i5}}{\|\boldsymbol{g}_i^u\|_2}\bigg)\mathfrak{m}(p-1,p)\bigg]\\
&\qquad+\mathbb{E}\big[O_{\prec}(\Psi^2)\mathfrak{m}(p-1,p)\big].
\end{aligned}
\tag{6.21}
$$

In addition, we also have the averaged analogue of (5.72):

$$
\begin{aligned}
&\frac{1}{N}\sum_i d_i\mathbb{E}\bigg[\frac{1}{\|\boldsymbol{g}_i^u\|_2}\bigg(\mathring{T}_{ii}-\frac{1}{N}\sum_k^{(i)}\frac{\partial(\hat{\boldsymbol{e}}_k^*G\hat{\boldsymbol{e}}_i)}{\partial g_{ik}^u}\bigg)\tau_1(\widetilde{B}G)\mathfrak{m}(p-1,p)\bigg]\\
&\quad=\frac{1}{N^2}\sum_i\sum_k^{(i)}d_i\mathbb{E}\bigg[\frac{\partial\|\boldsymbol{g}_i^u\|_2^{-2}}{\partial g_{ik}^u}\hat{\boldsymbol{e}}_k^*G\hat{\boldsymbol{e}}_i\tau_1(\widetilde{B}G)\mathfrak{m}(p-1,p)\bigg]\\
&\qquad+\frac{1}{N^2}\sum_i\sum_k^{(i)}d_i\mathbb{E}\bigg[\frac{\partial\tau_1(\widetilde{B}G)}{\partial g_{ik}^u}\frac{1}{\|\boldsymbol{g}_i^u\|_2^2}\hat{\boldsymbol{e}}_k^*G\hat{\boldsymbol{e}}_i\mathfrak{m}(p-1,p)\bigg]
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{p-1}{N^2}\sum_i\sum_k^{(i)} d_i\mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i^u\|_2^2}\hat{\boldsymbol{e}}_k^* G\hat{\boldsymbol{e}}_i\tau_1(\widetilde{B}G)\\
&\times\Big(\frac{1}{N}\sum_j d_j\frac{\partial\mathcal{Q}_{jj}}{\partial g_{ik}^u}\Big)\mathfrak{m}(p-2,p)\Big]\\
&+\frac{p}{N^2}\sum_i\sum_k^{(i)} d_i\mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i^u\|_2^2}\hat{\boldsymbol{e}}_k^* G\hat{\boldsymbol{e}}_i\tau_1(\widetilde{B}G)\\
&\times\Big(\frac{1}{N}\sum_j \bar{d}_j\frac{\partial\overline{\mathcal{Q}_{jj}}}{\partial g_{ik}^u}\Big)\mathfrak{m}(p-1,p-1)\Big].
\end{aligned}
\tag{6.22}
$$

Hence, to show (6.20), it suffices to estimate the second to the fifth terms on the right-hand side of (6.21), and the terms on the right-hand side of (6.22). First we notice that

$$
\varepsilon_{i4}=O_{\prec}(\Psi^2),
\tag{6.23}
$$

which can be seen from (5.25), (5.29) and the facts $|\|\boldsymbol{g}_i^a\|_2^2-1|\prec\frac{1}{\sqrt{N}}$ and $|h_{ii}^u|\prec\frac{1}{\sqrt{N}}$. All the other desired estimates can be derived from the following lemma.

LEMMA 6.3. *Suppose that the assumptions in Theorem 4.3 hold. Let $\eta_{\mathrm{M}}>0$ be any (large) constant and $\gamma>0$ in (5.1) be any (small) constant. Let $\hat{d}_1,\dots,\hat{d}_N\in\mathbb{C}$ be deterministic numbers with the bound $\max_i|\hat{d}_i|\lesssim 1$ and let $\tilde{d}_1,\dots,\tilde{d}_N\in\mathbb{C}$ be (possibly random) numbers with the bound $\max_i|\tilde{d}_i|\prec 1$ for all $i\in[\![1,N]\!]$. Let Q be any deterministic diagonal matrix satisfying $\|Q\|\le C$ and $X=\hat{I}$ or A, set $X_i=\hat{I}$ or $\widetilde{B}^{\langle i\rangle}$, and let*

$$
\boldsymbol{x}_i,\boldsymbol{y}_i=\begin{pmatrix}\mathring{\boldsymbol{g}}_i^u\\ \boldsymbol{0}\end{pmatrix}\quad\text{or}\quad\begin{pmatrix}\boldsymbol{0}\\ \mathring{\boldsymbol{g}}_i^v\end{pmatrix}.
$$

We have the estimates

$$
\begin{aligned}
&\frac{1}{N^2}\sum_{i=1}^N\sum_k^{(i)}\tilde{d}_i\frac{\partial\|\boldsymbol{g}_i^u\|_2^{-1}}{\partial g_{ik}^u}\hat{\boldsymbol{e}}_k^* X_i G\hat{\boldsymbol{e}}_i=O_{\prec}\Big(\frac{1}{N}\Big),\\
&\frac{1}{N^2}\sum_{i=1}^N\sum_k^{(i)}\tilde{d}_i\frac{\partial\operatorname{tr}QXG}{\partial g_{ik}^u}\hat{\boldsymbol{e}}_k^* X_i G\hat{\boldsymbol{e}}_i=O_{\prec}(\Psi^4),
\end{aligned}
\tag{6.24}
$$

uniformly on $\mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}},\eta_{\mathrm{M}})$. In addition, we also have

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N\hat{d}_i\mathbb{E}\big[(\boldsymbol{x}_i^*X_i\boldsymbol{y}_i-\mathbb{E}_i[\boldsymbol{x}_i^*X_i\boldsymbol{y}_i])\mathfrak{m}(p-1,p)\big]\\
&\quad=\mathbb{E}\big[O_{\prec}(\Psi^2)\mathfrak{m}(p-1,p)\big]+\mathbb{E}\big[O_{\prec}(\Psi^4)\mathfrak{m}(p-2,p)\big]\\
&\qquad+\mathbb{E}\big[O_{\prec}(\Psi^4)\mathfrak{m}(p-1,p-1)\big],
\end{aligned}
\tag{6.25}
$$

uniformly on $\mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$, *where* $\mathbb{E}_i$ *denotes the expectation with respect to* $\mathring{\boldsymbol{g}}_i^u$ *and* $\mathring{\boldsymbol{g}}_i^v$.

With Lemma 6.3, we can proceed to the proof of Theorem 6.2 as follows. First of all, for any diagonal matrix $Q = \operatorname{diag}(q_1, \dots, q_{2N})$, using the first estimate in (5.25), we have

$$\operatorname{tr} QG = \frac{1}{2N} \sum_{i=1}^{N} (q_i + q_{\hat{i}}) \frac{\omega_B(z)}{|\xi_i|^2 - (\omega_B(z))^2} + O_{\prec}(\Psi),$$

$$\operatorname{tr} QAG = \frac{1}{2N} \sum_{i=1}^{N} (q_i + q_{\hat{i}}) \frac{|\xi_i|^2}{|\xi_i|^2 - (\omega_B(z))^2} + O_{\prec}(\Psi).$$

Using the upper bound of ω_B and the lower bound of $\operatorname{Im} \omega_B$ in (A.4), we can see that

$$|\operatorname{tr} QXG| \prec 1, \tag{6.26}$$

for diagonal Q with $\|Q\| \le C$ and $X = \hat{I}$ or A. Note that all partial traces such as $\tau_1(G)$, $\tau_1(\widetilde{B}G)$ can be written as a linear combination of terms of the form $\operatorname{tr} QXG$ with the aid of the identities in (4.16), and thus for these partial traces we have

$$\tau_1(G) = O_{\prec}(1), \qquad \tau_1(\widetilde{B}G) = O_{\prec}(1).$$

These bounds together with the first estimate in (6.24), imply the desired estimates for the second term on the right-hand side of (6.21) and the first term on the right-hand side of (6.22).

Next notice that

$$\frac{1}{N} \sum_{j=1}^{N} d_j \mathcal{Q}_{jj} = \operatorname{tr}(D\widetilde{B}G)\tau_1(G) - \operatorname{tr}(DG)\tau_1(\widetilde{B}G) + \operatorname{tr}(DG)\Upsilon_1,$$

where we denoted the deterministic diagonal matrix $D := \operatorname{diag}(d_1, \dots, d_N) \oplus 0$, with 0 the $N \times N$ zero matrix. In addition, using (4.16), we can see that $\frac{1}{N}\sum_j d_j \mathcal{Q}_{jj}$ is a polynomial of $\frac{1}{N}\sum_j d_j T_{jj}$ and the terms of the form $\operatorname{tr} QXG$ for some diagonal Q with $\|Q\| \le C$ and $X = \hat{I}$ or A. Here, we also used the fact that $\tau_a(\mathcal{D}) = \operatorname{tr}(\hat{I}_a \mathcal{D})$ for any $\mathcal{D} \in M_{2N}(\mathbb{C})$ and $a = 1, 2$, where $\hat{I}_a$ is defined in (4.25). Then the last two estimates in (6.24), (6.26), together with the chain rule, imply that

$$\frac{1}{N^3} \sum_{i=1}^{N} \sum_{k}^{(i)} \tilde{d}_i \hat{\boldsymbol{e}}_k^* X_i G \hat{\boldsymbol{e}}_i \sum_{j=1}^{N} d_j \frac{\partial \mathcal{Q}_{jj}}{\partial g_{ik}^u} = O_{\prec}(\Psi^4). \tag{6.27}$$

Similarly, we can prove the same bound if we replace $\mathcal{Q}_{jj}$'s by $\overline{\mathcal{Q}}_{jj}$'s. Hence, the desired estimates for the third to the fifth terms on the right-hand side of (6.21), and the last three terms on the right-hand side of (6.22) can be obtained from the second estimate in (6.24).

Hence, what remains is to estimate the sixth term in (6.21). First, according to (6.23), we can neglect ε_{i4}. Then we recall the definition of ε_{i1} from (5.43). Using the estimates of $G_{\hat{i}i}$ and $T_{\hat{i}i}$ from the first and the third inequalities in (5.25), and the estimates

$$\ell_i^v = 1 + O_\prec\Big(\frac{1}{\sqrt{N}}\Big), \qquad |h_{ii}^u| \prec \frac{1}{\sqrt{N}}, \qquad |(\boldsymbol{k}_i^u)^* \widetilde{B}^{\langle i\rangle} \boldsymbol{k}_i^v| \prec \frac{1}{\sqrt{N}},$$

we see that

$$\begin{aligned} \varepsilon_{i1} &= \frac{\bar{\xi}_i}{|\xi_i|^2 - \omega_B^2} (\boldsymbol{k}_i^u)^* \widetilde{B}^{\langle i\rangle} \boldsymbol{k}_i^v + O_\prec(\Psi^2) \\ &= \frac{\bar{\xi}_i}{|\xi_i|^2 - \omega_B^2} (\boldsymbol{\ell}_i^u)^* \widetilde{B}^{\langle i\rangle} \boldsymbol{\ell}_i^v + O_\prec(\Psi^2), \end{aligned} \tag{6.28}$$

where we introduced the notation:

$$\boldsymbol{\ell}_i^u := \begin{pmatrix} \mathring{\boldsymbol{g}}_i^u \\ \mathbf{0} \end{pmatrix}, \qquad \boldsymbol{\ell}_i^v := \begin{pmatrix} \mathbf{0} \\ \mathring{\boldsymbol{g}}_i^v \end{pmatrix}.$$

Then recall the definition of ε_{i5} from (5.69). Applying the estimate of G_{ii} from the first inequality in (5.25), and the second formula in (6.2, and the fact $\|\boldsymbol{g}_i^u\|_2^2 = \|\boldsymbol{\ell}_i^u\|_2^2 + O_\prec(\frac{1}{N}) = 1 + O_\prec(\frac{1}{\sqrt{N}})$, we also have

$$\varepsilon_{i5} = \frac{(z - \omega_B) m_{\mu_A \boxplus \mu_B} \omega_B}{|\xi_i|^2 - \omega_B^2} (\|\boldsymbol{\ell}_i^u\|_2^2 - 1) + O_\prec(\Psi^2). \tag{6.29}$$

Note that both of the first terms on the right-hand side of (6.28) and (6.29) are of the form $\hat{d}_i(\boldsymbol{x}_i^* X_i \boldsymbol{y}_i - \mathbb{E}_i[\boldsymbol{x}_i^* X_i \boldsymbol{y}_i])$ for some deterministic $\hat{d}_i$ with $|\hat{d}_i| \lesssim 1$. Hence, using (6.25), we get the desired bound for the sixth term of (6.21). This completes the proof of Theorem 6.2 up to the proof of Lemma 6.3. □

PROOF OF LEMMA 6.3. The first estimate in (6.24) follows directly from the first estimate in (5.75). The second estimate of (6.24) is a weighted average of the last estimate in (5.75).

Hence, what remains is to prove (6.25). We only show the details for the case $\boldsymbol{x}_i = \boldsymbol{\ell}_i^u$ and $\boldsymbol{y}_i = \boldsymbol{\ell}_i^v$. The others are similar. Notice that in this case, $\mathbb{E}_i[\boldsymbol{x}_i^* X_i \boldsymbol{y}_i] = 0$. Using the integration by parts formula (5.44) again, we have

$$\begin{aligned} &\frac{1}{N} \sum_i \hat{d}_i \mathbb{E}\big[(\boldsymbol{\ell}_i^u)^* X_i \boldsymbol{\ell}_i^v \mathfrak{m}(p-1, p)\big] \\ &\quad = \frac{1}{N} \sum_i \sum_k^{(i)} \hat{d}_i \mathbb{E}\big[\bar{g}_{ik}^u \hat{\boldsymbol{e}}_k^* X_i \boldsymbol{\ell}_i^v \mathfrak{m}(p-1, p)\big] \end{aligned}$$

$$= \frac{p-1}{N^2} \sum_i \sum_k^{(i)} \hat{d}_i \mathbb{E}\Big[\hat{\boldsymbol{e}}_k^* X_i \boldsymbol{\ell}_i^v \frac{1}{N} \sum_j d_j \frac{\partial \mathcal{Q}_{jj}}{\partial g_{ik}^u} \mathfrak{m}(p-2, p)\Big]$$

$$+ \frac{p}{N^2} \sum_i \sum_k^{(i)} \hat{d}_i \mathbb{E}\Big[\hat{\boldsymbol{e}}_k^* X_i \boldsymbol{\ell}_i^v \frac{1}{N} \sum_j \bar{d}_j \frac{\partial \overline{\mathcal{Q}}_{jj}}{\partial g_{ik}^u} \mathfrak{m}(p-1, p-1)\Big].$$

Hence, it suffices to show

$$\Big| \frac{1}{N^3} \sum_i \sum_k^{(i)} \hat{d}_i \hat{\boldsymbol{e}}_k^* X_i \boldsymbol{\ell}_i^v \sum_j d_j \frac{\partial \mathcal{Q}_{jj}}{\partial g_{ik}^u} \Big| \prec \Psi^4 \tag{6.30}$$

and its complex conjugate analogue. The proof of (6.30) is nearly the same as that of (6.27). Hence, we omit it. Therefore, we completed the proof of Lemma 6.3. □

Acknowledgments. Part of this work was accomplished when Z.-G. B. and K. S. were working at IST Austria with the support of ERC Advanced Grant RANMAT No. 338804. Support and hospitality are gratefully acknowledged.

SUPPLEMENTARY MATERIAL

Supplement to "Local single ring theorem on optimal scale" (DOI: 10.1214/18-AOP1284SUPP; .pdf). We establish the proof of Theorem 2.2 and the proof of Theorem 4.3 for large η. Moreover, in the appendices at the end we collect some auxiliary information.

REFERENCES

[1] ANDERSON, G. W., GUIONNET, A. and ZEITOUNI, O. (2010). *An Introduction to Random Matrices*. *Cambridge Studies in Advanced Mathematics* **118**. Cambridge Univ. Press, Cambridge. MR2760897

[2] BAO, Z., ERDŐS, L. and SCHNELLI, K. (2016). Local stability of the free additive convolution. *J. Funct. Anal.* **271** 672–719. MR3506962

[3] BAO, Z., ERDŐS, L. and SCHNELLI, K. (2017). Local law of addition of random matrices on optimal scale. *Comm. Math. Phys.* **349** 947–990. MR3602820

[4] BAO, Z., ERDŐS, L. and SCHNELLI, K. (2017). Convergence rate for spectral distribution of addition of random matrices. *Adv. Math.* **319** 251–291. MR3695875

[5] BAO, Z. G., ERDŐS, L. and SCHNELLI, K. (2019). Supplement to "Local single ring theorem on optimal scale." DOI:10.1214/18-AOP1284SUPP.

[6] BAO, Z. G., ERDŐS, L. and SCHNELLI, K. (2017). Spectral rigidity for addition of random matrices at the regular edge. Preprint. Available at arXiv:1708.01597.

[7] BELINSCHI, S. T. (2006). A note on regularity for free convolutions. *Ann. Inst. Henri Poincaré Probab. Stat.* **42** 635–648. MR2259979

[8] BELINSCHI, S. T. (2008). The Lebesgue decomposition of the free additive convolution of two probability distributions. *Probab. Theory Related Fields* **142** 125–150. MR2413268

[9] BELINSCHI, S. T. and BERCOVICI, H. (2007). A new approach to subordination results in free probability. *J. Anal. Math.* **101** 357–365. MR2346550

[10] BENAYCH-GEORGES, F. (2015). Exponential bounds for the support convergence in the single ring theorem. *J. Funct. Anal.* **268** 3492–3507. MR3336731
[11] BENAYCH-GEORGES, F. (2017). Local single ring theorem. *Ann. Probab.* **45** 3850–3885. MR3729617
[12] BERCOVICI, H. and VOICULESCU, D. (1993). Free convolution of measures with unbounded support. *Indiana Univ. Math. J.* **42** 733–773. MR1254116
[13] BIANE, P. (1998). Processes with free increments. *Math. Z.* **227** 143–174. MR1605393
[14] BORDENAVE, C. and CHAFAÏ, D. (2012). Around the circular law. *Probab. Surv.* **9** 1–89. MR2908617
[15] BOURGADE, P., YAU, H.-T. and YIN, J. (2014). Local circular law for random matrices. *Probab. Theory Related Fields* **159** 545–595. MR3230002
[16] BOURGADE, P., YAU, H.-T. and YIN, J. (2014). The local circular law II: The edge case. *Probab. Theory Related Fields* **159** 619–660. MR3230004
[17] CHISTYAKOV, G. P. and GÖTZE, F. (2011). The arithmetic of distributions in free probability theory. *Cent. Eur. J. Math.* **9** 997–1050. MR2824443
[18] DIACONIS, P. and SHAHSHAHANI, M. (1987). The subgroup algorithm for generating uniform random variables. *Probab. Engrg. Inform. Sci.* **1** 15–32.
[19] ERDŐS, L. (2014). Random matrices, log-gases and Hölder regularity. In *Proceedings of the International Congress of Mathematicians—Seoul* 2014. *Vol. III* 213–236. Kyung Moon Sa, Seoul. MR3729025
[20] ERDŐS, L., KNOWLES, A. and YAU, H.-T. (2013). Averaging fluctuations in resolvents of random band matrices. *Ann. Henri Poincaré* **14** 1837–1926. MR3119922
[21] ERDŐS, L., SCHLEIN, B. and YAU, H.-T. (2009). Local semicircle law and complete delocalization for Wigner random matrices. *Comm. Math. Phys.* **287** 641–655. MR2481753
[22] ERDŐS, L., YAU, H.-T. and YIN, J. (2011). Universality for generalized Wigner matrices with Bernoulli distribution. *J. Comb.* **2** 15–81. MR2847916
[23] ERDŐS, L., YAU, H.-T. and YIN, J. (2012). Bulk universality for generalized Wigner matrices. *Probab. Theory Related Fields* **154** 341–407. MR2981427
[24] FEINBERG, J. and ZEE, A. (1997). Non-Gaussian non-Hermitian random matrix theory: Phase transition and addition formalism. *Nuclear Phys. B* **501** 643–669. MR1477381
[25] GIRKO, V. L. (1984). The circular law. *Teor. Veroyatn. Primen.* **29** 669–679. MR0773436
[26] GUIONNET, A., KRISHNAPUR, M. and ZEITOUNI, O. (2011). The single ring theorem. *Ann. of Math.* (2) **174** 1189–1217. MR2831116
[27] GUIONNET, A. and ZEITOUNI, O. (2012). Support convergence in the single ring theorem. *Probab. Theory Related Fields* **154** 661–675. MR3000558
[28] HAAGERUP, U. and LARSEN, F. (2000). Brown's spectral distribution measure for R-diagonal elements in finite von Neumann algebras. *J. Funct. Anal.* **176** 331–367. MR1784419
[29] KARGIN, V. (2015). Subordination for the sum of two random matrices. *Ann. Probab.* **43** 2119–2150. MR3353823
[30] LEE, J. O. and SCHNELLI, K. (2018). Local law and Tracy–Widom limit for sparse random matrices. *Probab. Theory Related Fields* **171** 543–616. MR3800840
[31] MEZZADRI, F. (2007). How to generate random matrices from the classical compact groups. *Notices Amer. Math. Soc.* **54** 592–604. MR2311982
[32] PASTUR, L. and VASILCHUK, V. (2000). On the law of addition of random matrices. *Comm. Math. Phys.* **214** 249–286. MR1796022
[33] RUDELSON, M. and VERSHYNIN, R. (2014). Invertibility of random matrices: Unitary and orthogonal perturbations. *J. Amer. Math. Soc.* **27** 293–338. MR3164983
[34] TAO, T. and VU, V. (2010). Random matrices: Universality of ESDs and the circular law. *Ann. Probab.* **38** 2023–2065. MR2722794
[35] TAO, T. and VU, V. (2015). Random matrices: Universality of local spectral statistics of non-Hermitian matrices. *Ann. Probab.* **43** 782–874. MR3306005

[36] VOICULESCU, D. (1993). The analogues of entropy and of Fisher's information measure in free probability theory. I. *Comm. Math. Phys.* **155** 71–92. MR1228526

[37] YIN, J. (2014). The local circular law III: General case. *Probab. Theory Related Fields* **160** 679–732. MR3278919

Z. BAO
DEPARTMENT OF MATHEMATICS
HONG KONG UNIVERSITY OF SCIENCE
AND TECHNOLOGY
CLEAR WATER BAY
KOWLOON
HONG KONG
E-MAIL: mazgbao@ust.hk

L. ERDŐS
INSTITUTE OF SCIENCE AND TECHNOLOGY
AUSTRIA
AM CAMPUS 1
3400 KLOSTERNEUBURG
AUSTRIA
E-MAIL: lerdos@ist.ac.at

K. SCHNELLI
DEPARTMENT OF MATHEMATICS
KTH ROYAL INSTITUTE OF TECHNOLOGY
LINDSTEDTSVÄGEN 25
100 44 STOCKHOLM
SWEDEN
E-MAIL: schnelli@kth.se