Journal of Functional Analysis 279 (2020) 108639

Contents lists available at ScienceDirect

Journal of Functional Analysis

www.elsevier.com/locate/jfa

Spectral rigidity for addition of random matrices at the regular edge

Zhigang Bao [a,*,1], László Erdős [b,2], Kevin Schnelli [c,3]

[a] *HKUST, Hong Kong*
[b] *IST, Austria*
[c] *KTH Royal Institute of Technology, Sweden*

A R T I C L E   I N F O

Article history:
Received 31 August 2017
Accepted 7 May 2020
Available online 14 May 2020
Communicated by S. Vaes

MSC:
46L54
60B20

Keywords:
Random matrices
Local eigenvalue density
Free convolution
Spectral edge

A B S T R A C T

We consider the sum of two large Hermitian matrices A and B with a Haar unitary conjugation bringing them into a general relative position. We prove that the eigenvalue density on the scale slightly above the local eigenvalue spacing is asymptotically given by the free additive convolution of the laws of A and B as the dimension of the matrix increases. This implies optimal rigidity of the eigenvalues and optimal rate of convergence in Voiculescu's theorem. Our previous works [4,5] established these results in the bulk spectrum, the current paper completely settles the problem at the spectral edges provided they have the typical square-root behavior. The key element of our proof is to compensate the deterioration of the stability of the subordination equations by sharp error estimates that properly account for the local density near the edge. Our results also hold if the Haar unitary matrix is replaced by the Haar orthogonal matrix.

* Corresponding author.
E-mail addresses: mazgbao@ust.hk (Z. Bao), lerdos@ist.ac.at (L. Erdős), schnelli@kth.se (K. Schnelli).

[1] Partially supported by Hong Kong RGC grants ECS 26301517 and GRF 16301519.
[2] Partially supported by ERC Advanced Grant RANMAT No. 338804.
[3] Partially supported by the Göran Gustafsson Foundation and the Swedish Research Council grant VR-2017-05195.

https://doi.org/10.1016/j.jfa.2020.108639
0022-1236/

1. Introduction

The pioneering work of Voiculescu [28] identified the eigenvalue density of the sum of two Hermitian $N \times N$ matrices A and B in a general relative position as the free additive convolution of the eigenvalue densities μ_A and μ_B of A and B. The primary example for general relative positions is asymptotic freeness that can be generated by conjugation via a Haar distributed unitary matrix. In fact, under some mild regularity condition on μ_A and μ_B, *local laws* also hold, asserting that the empirical eigenvalue density of the sum converges on small scales as well. The optimal precision in such local law pins down the location of individual eigenvalues with an error bar that is just slightly above the local eigenvalue spacing. With an optimal error term, it identifies the speed of convergence of order $N^{-1+\epsilon}$ in Voiculescu's limit theorem.

After several gradual improvements on the precision in [20,21,3], the local law on the optimal $N^{-1+\epsilon}$ scale was established in [4] and the optimal convergence speed was obtained in [5]. All these results were, however, restricted to the *regular bulk* spectrum, *i.e.*, to the spectral regime where the density of the free convolution is non-vanishing and bounded from above. In particular, the regime of the spectral edges were not covered. Under mild conditions on the limiting eigenvalue densities of A and B, the free convolution density always vanishes as the square-root function near the edges of its support. We call such type of edges *regular*. We remark that the regular edge is typical in many random matrix models, for instance, the semicircle law; *i.e.*, the limiting density for Wigner matrices.

Near the edges the eigenvalues are sparser hence they fluctuate more; naively, the extreme eigenvalues might be prone to very large fluctuations due to the room available to them on the opposite side of the support. Nevertheless, for Wigner matrices and many related ensembles with independent or weakly dependent entries it has been shown that the eigenvalue fluctuation does not exceed its natural threshold, the local spacing, even at the edge; see *e.g.*, [18,22,2] and references therein. In general, it implies a very strong concentration of the empirical measure. For the smallest and largest eigenvalues it means a fluctuation of order $N^{-2/3}$. In fact, the precise fluctuation is universal and it follows the Tracy–Widom distribution; see *e.g.*, [26,12,23] for proofs in various models.

In this paper we present a comprehensive edge local law on optimal scale and with optimal precision for the ensemble $A+UBU^*$ where U is Haar unitary. We assume that the laws of A and B are close to continuous limiting profiles μ_α and μ_β with a single interval support and power law behavior at the edge with exponent less than one. We prove that the free convolution $\mu_\alpha \boxplus \mu_\beta$ has a square root singularity at its edge and $\mu_A \boxplus \mu_B$ closely trails this behavior. Furthermore, we establish that the eigenvalues of $A+UBU^*$ follow $\mu_A \boxplus \mu_B$ down to the scale of the local spacing, uniformly throughout the spectrum. In particular, we show that the extreme eigenvalues are in the optimal $N^{-\frac{2}{3}+\varepsilon}$ vicinity of the deterministic spectral edges. Previously, similar results were only known with $o(1)$ precision, see [15] for instance. We expect that Tracy–Widom law holds at the regular edge of our additive model. Very recently, bulk universality has been demonstrated in [13].

Our analysis also implies optimal rate of convergence for Voiculescu's global law for free convolution densities with the typical square root edges.

The result demonstrates that the Haar randomness in the additive model has a similarly strong concentration of the empirical density as already proved for the Wigner ensemble earlier. In fact, the additive model is only the simplest prototype of a large family of models involving polynomials of Haar unitaries and deterministic matrices; other examples include the ensemble in the single ring theorem [19,6]. The technique developed in the current paper can potentially handle square root edges in more complicated ensembles where the main source of randomness is the Haar unitaries.

After the statement of the main result and the introduction of a few basic quantities, we show in Section 3 that $\mu_\alpha \boxplus \mu_\beta$ has under suitable conditions a square root singularity at the lowest edge and we establish stability properties of the subordination equations around that edge. In Section 4 an informal outline of the proof that explains the main difficulties stemming from the edge in contrast to the related analysis in the bulk. Here we highlight only the key point. A typical proof of the local laws has two parts: (i) stability analysis of a deterministic (Dyson) equation for the limiting eigenvalue distribution, and (ii) proof that the empirical density approximately satisfies the Dyson equation and estimate the error. Given these two inputs, the local law follows by simply inverting the Dyson equation. For our model the Dyson equation is actually the pair of the *subordination equations*, that define the free convolution. Near the spectral edge, the subordination equations become unstable. A similar phenomenon is well known for the Dyson equation of Wigner type models, but it has not yet been analyzed for the subordination equations. This instability can only be compensated by a very accurate estimate on the approximation error; a formidable task given the complexity of the analogous error estimates in the bulk [5]. Already the bulk analysis required carefully selected counter terms and weights in the fluctuation averaging mechanisms before recursive moment estimates could be started. All these ideas are used at the edge, even up to higher order, but they still fall short of the necessary precision. The key novelty is to identify a very specific linear combination of two basic fluctuating quantities with a fluctuation smaller than those of its constituencies, indicating a very special strong correlation between them.

Notation: The symbols $O(\,\cdot\,)$ and $o(\,\cdot\,)$ stand for the standard big-O and little-o notation. We use c and C to denote positive finite constants that do not depend on the matrix size 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 vector $\boldsymbol{v} \in \mathbb{C}^N$, we use $\|\boldsymbol{v}\|$ to denote its Euclidean norm. For $A \in M_N(\mathbb{C})$, we denote by $\|A\|$ its operator norm and by $\|A\|_2$ its Hilbert-Schmidt norm. We use $\operatorname{tr} A = \frac{1}{N}\sum_i A_{ii}$ to denote the normalized trace of an $N \times N$ matrix $A = (A_{ij})_{N,N}$.

Let $\boldsymbol{g} = (g_1, \ldots, g_N)$ be a real or complex Gaussian vector. We write $\boldsymbol{g} \sim \mathcal{N}_{\mathbb{R}}(0, \sigma^2 I_N)$ if $g_1, \ldots, g_N$ 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 $g_1, \ldots, g_N$ 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.

For two possibly N-dependent numbers $a, b \in \mathbb{C}$, we write $a \sim b$ if there is a (large) positive constant $C > 1$ such that $C^{-1}|a| \leq |b| \leq C|a|$. We use double brackets to denote index sets, *i.e.*, for $n_1, n_2 \in \mathbb{R}$, $[\![n_1, n_2]\!] := [n_1, n_2] \cap \mathbb{Z}$, and denote by $\mathbb{C}^+$ the upper complex half-plane, *i.e.*, $\mathbb{C}^+ := \{z \in \mathbb{C} : \operatorname{Im} z > 0\}$.

2. Definition of the model and main results

2.1. Model and assumptions

Let $A \equiv A_N = \operatorname{diag}(a_1, \ldots, a_N)$ and $B \equiv B_N = \operatorname{diag}(b_1, \ldots, b_N)$ be two deterministic real diagonal matrices in $M_N(\mathbb{C})$. Let $U \equiv U_N$ be a random unitary matrix which is Haar distributed on $\mathcal{U}(N)$, where $\mathcal{U}(N)$ is the N-dimensional unitary group. We study the following random Hermitian matrix

$$H \equiv H_N := A + UBU^*. \tag{2.1}$$

More specifically, we study the eigenvalues of H, denoted by $\lambda_1 \leq \ldots \leq \lambda_N$. Throughout the paper, we are mainly working in the vicinity of the bottom of the spectrum. The discussion for the top of the spectrum is analogous. Let μ_A, μ_B and μ_H be the empirical eigenvalue distributions of A, B, and H, *i.e.*,

$$\mu_A := \frac{1}{N} \sum_{i=1}^{N} \delta_{a_i}, \qquad \mu_B := \frac{1}{N} \sum_{i=1}^{N} \delta_{b_i}, \qquad \mu_H := \frac{1}{N} \sum_{i=1}^{N} \delta_{\lambda_i}.$$

For any probability measure μ on the real line, its Stieltjes transform is defined as

$$m_\mu(z) := \int_{\mathbb{R}} \frac{1}{x - z} \, \mathrm{d}\mu(x), \qquad z \in \mathbb{C}^+,$$

where z is called the *spectral parameter*. Throughout the paper, we write $z = E + \mathrm{i}\eta$, *i.e.*, $E = \operatorname{Re} z$, $\operatorname{Im} z = \eta$.

In this paper, we assume that there are two N-independent absolutely continuous probability measures μ_α and μ_β with continuous density functions ρ_α and ρ_β, respectively, such that the following assumptions, Assumptions 2.1 and 2.2, are satisfied. The first one discusses some qualitative properties of μ_α and μ_β, while the second one demands that μ_A and μ_B are close to μ_α and μ_β, respectively.

Assumption 2.1. We assume the following:

(i) Both density functions ρ_α and ρ_β have single non-empty interval supports, $[E_-^\alpha, E_+^\alpha]$ and $[E_-^\beta, E_+^\beta]$, respectively, and ρ_α and ρ_β are strictly positive in the interior of their supports.

(ii) In a small δ-neighborhood of the lower edges of the supports, these measures have a power law behavior, namely, there is a (small) constant $\delta > 0$ and exponents $-1 < t_-^\alpha, t_-^\beta < 1$ such that

$$C^{-1} \leq \frac{\rho_\alpha(x)}{(x - E_-^\alpha)^{t_-^\alpha}} \leq C\,, \qquad \forall x \in [E_-^\alpha, E_-^\alpha + \delta]\,,$$

$$C^{-1} \leq \frac{\rho_\beta(x)}{(x - E_-^\beta)^{t_-^\beta}} \leq C\,, \qquad \forall x \in [E_-^\beta, E_-^\beta + \delta]\,,$$

hold for some positive constant $C > 1$.

(iii) We assume that at least one of the following two bounds holds

$$\sup_{z \in \mathbb{C}^+} |m_{\mu_\alpha}(z)| \leq C\,, \qquad \sup_{z \in \mathbb{C}^+} |m_{\mu_\beta}(z)| \leq C\,, \tag{2.2}$$

for some positive constant C.

Assumption 2.2. We assume the following:

(iv) For the Lévy-distances d_L, we have that

$$\boldsymbol{d} := d_L(\mu_A, \mu_\alpha) + d_L(\mu_B, \mu_\beta) \leq N^{-1+\epsilon}\,, \tag{2.3}$$

for any constant $\epsilon > 0$ when N is sufficiently large.

(v) For the lower edges, we have

$$\inf \operatorname{supp} \mu_A \geq E_-^\alpha - \delta\,, \qquad \inf \operatorname{supp} \mu_B \geq E_-^\beta - \delta\,, \tag{2.4}$$

for any constant $\delta > 0$ when N is sufficiently large.

(vi) For the upper edges, we assume that there is a constant C such that

$$\sup \operatorname{supp} \mu_A \leq C\,, \qquad \sup \operatorname{supp} \mu_B \leq C\,. \tag{2.5}$$

A direct consequence of (v) and (vi) above is that there is a constant C' such that $\|A\|, \|B\| \leq C'$.

Since [28], it is well known now that μ_H can be weakly approximated by a deterministic probability measure, called the free additive convolution of μ_A and μ_B. Here we briefly introduce some notations concerning the free additive convolution, which will be necessary to state our main results.

For a probability measure μ on $\mathbb{R}$, we denote by F_μ its negative reciprocal Stieltjes transform, *i.e.*,

$$F_\mu(z) := -\frac{1}{m_\mu(z)}\,, \qquad\qquad z \in \mathbb{C}^+\,. \tag{2.6}$$

Note that $F_\mu : \mathbb{C}^+ \to \mathbb{C}^+$ is analytic and satisfies

$$\lim_{\eta \nearrow \infty} \frac{F_\mu(\mathrm{i}\eta)}{\mathrm{i}\eta} = 1\,. \tag{2.7}$$

Conversely, if $F : \mathbb{C}^+ \to \mathbb{C}^+$ is an analytic function with $\lim_{\eta\nearrow\infty} F(\mathrm{i}\eta)/\mathrm{i}\eta = 1$, then F is the negative reciprocal Stieltjes transform of a probability measure μ, *i.e.*, $F(z) = F_\mu(z)$, for all $z \in \mathbb{C}^+$; see *e.g.*, [1].

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

Proposition 2.3 *(Theorem 4.1 in [9], Theorem 2.1 in [14]). 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

$$\lim_{\eta\nearrow\infty} \frac{\omega_1(\mathrm{i}\eta)}{\mathrm{i}\eta} = \lim_{\eta\nearrow\infty} \frac{\omega_2(\mathrm{i}\eta)}{\mathrm{i}\eta} = 1\,; \tag{2.8}$$

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

$$F_{\mu_1}(\omega_2(z)) = F_{\mu_2}(\omega_1(z))\,, \qquad\qquad \omega_1(z) + \omega_2(z) - z = F_{\mu_1}(\omega_2(z))\,. \tag{2.9}$$

The analytic function $F : \mathbb{C}^+ \to \mathbb{C}^+$ defined by

$$F(z) := F_{\mu_1}(\omega_2(z)) = F_{\mu_2}(\omega_1(z))\,, \tag{2.10}$$

is, in virtue of (2.8), 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*. The subordination phenomenon for the addition of freely independent non-commutative random variables was first noted by Voiculescu [29] in a generic situation and extended to full generality by Biane [11].

Choosing $(\mu_1, \mu_2) = (\mu_\alpha, \mu_\beta)$ in Proposition 2.3, we denote the associated subordination functions ω_1 and ω_2 by ω_α and ω_β, respectively. Analogously, for the choice $(\mu_1, \mu_2) = (\mu_A, \mu_B)$, we denote by ω_A and ω_B the associated subordination functions. With the above notations, we obtain from (2.9) and (2.10) the following subordination equations

$$m_{\mu_A}(\omega_B(z)) = m_{\mu_B}(\omega_A(z)) = m_{\mu_A \boxplus \mu_B}(z),$$

$$\omega_A(z) + \omega_B(z) - z = -\frac{1}{m_{\mu_A \boxplus \mu_B}(z)}. \tag{2.11}$$

The same system of equations holds if we replace the subscripts (A, B) by (α, β).

We denote the lower and upper edges of the support of $\mu_\alpha \boxplus \mu_\beta$ by

$$E_- := \inf \operatorname{supp} \mu_\alpha \boxplus \mu_\beta\,, \qquad E_+ := \sup \operatorname{supp} \mu_\alpha \boxplus \mu_\beta\,. \tag{2.12}$$

In Section 3, we establish various qualitative properties of $\mu_\alpha \boxplus \mu_\beta$ and of $\mu_A \boxplus \mu_B$. In particular, under Assumption 2.1, we show that $\mu_\alpha \boxplus \mu_\beta$ has a square-root decay at the lower edge, see (3.63).

2.2. Main results

To state our results, we introduce some more terminology. We denote the Green function or resolvent of H and its normalized trace by

$$G(z) \equiv G_H(z) := \frac{1}{H - z}\,, \qquad m_H(z) := \operatorname{tr} G(z) = \frac{1}{N} \sum_{i=1}^N G_{ii}(z)\,, \qquad z \in \mathbb{C}^+.$$

Observe that $m_H(z)$ is also the Stieltjes transform of μ_H, *i.e.*,

$$m_H(z) = \int_{\mathbb{R}} \frac{1}{x - z} \mathrm{d}\mu_H(x) = \frac{1}{N} \sum_{i=1}^N \frac{1}{\lambda_i - z}, \qquad z \in \mathbb{C}^+.$$

We further set

$$\mathcal{K} := \|A\| + \|B\| + 1\,. \tag{2.13}$$

Moreover, for any spectral parameter $z = E + \mathrm{i}\eta \in \mathbb{C}^+$, we let

$$\kappa \equiv \kappa(z) := \min\{|E - E_-|, |E - E_+|\}\,, \tag{2.14}$$

with $E_\pm$ given in (2.12). We then introduce the following domain of the spectral parameter z: For any $0 < a \le b$ and $0 < \tau < \frac{E_+ - E_-}{2}$,

$$\mathcal{D}_\tau(a, b) := \{z = E + \mathrm{i}\eta \in \mathbb{C}^+ : -\mathcal{K} \le E \le E_- + \tau, \quad a \le \eta \le b\}. \tag{2.15}$$

For any (small) positive constant $\gamma > 0$, we set

$$\eta_{\mathrm{m}} := N^{-1+\gamma}.$$

Let $\eta_{\mathrm{M}} > 1$ be some sufficiently large constant. In the rest of the paper, we will mainly work in the regime $z \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$ with sufficiently small constant $\tau > 0$. In particular, we usually have $\eta_{\mathrm{m}} \leq \eta \leq \eta_{\mathrm{M}}$.

We also need the following definition on high-probability estimates from [17]. In Appendix A we collect some of its properties.

Definition 2.4. Let $\mathcal{X} \equiv \mathcal{X}^{(N)}$ and $\mathcal{Y} \equiv \mathcal{Y}^{(N)}$ be two sequences of nonnegative random variables. We say that $\mathcal{Y}$ stochastically dominates $\mathcal{X}$ if, for all (small) $\epsilon > 0$ and (large) $D > 0$,

$$\mathbb{P}\big(\mathcal{X}^{(N)} > N^{\epsilon}\mathcal{Y}^{(N)}\big) \leq N^{-D}, \tag{2.16}$$

for sufficiently large $N \geq N_0(\epsilon, D)$, and we write $\mathcal{X} \prec \mathcal{Y}$ or $\mathcal{X} = O_\prec(\mathcal{Y})$. 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.

With these definitions and notations, we now state our main result.

Theorem 2.5 *(Local law at the regular edge). Suppose that Assumptions 2.1 and 2.2 hold. Let $\tau > 0$ be a sufficiently small constant and fix any (small) constants $\gamma > 0$ and $\varepsilon > 0$. Let $d_1, \ldots, d_N \in \mathbb{C}$ be any deterministic complex number satisfying*

$$\max_{i \in [\![1,N]\!]} |d_i| \leq 1.$$

Then

$$\Big|\frac{1}{N}\sum_{i=1}^{N} d_i\Big(G_{ii}(z) - \frac{1}{a_i - \omega_B(z)}\Big)\Big| \prec \frac{1}{N\eta} \tag{2.17}$$

holds uniformly on $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$ with $\eta_{\mathrm{m}} = N^{-1+\gamma}$ and any constant $\eta_{\mathrm{M}} > 0$. In particular, choosing $d_i = 1$ for all $i \in [\![1, N]\!]$, we have the estimate

$$\Big|m_H(z) - m_{\mu_A \boxplus \mu_B}(z)\Big| \prec \frac{1}{N\eta}, \tag{2.18}$$

uniformly on $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$. Moreover, we have the improved estimate

$$\Big|m_H(z) - m_{\mu_A \boxplus \mu_B}(z)\Big| \prec \frac{1}{N(\kappa + \eta)}, \tag{2.19}$$

uniformly for all $z = E + \mathrm{i}\eta \in \mathcal{D}_\tau(0, \eta_{\mathrm{M}})$ with $E \leq E_- - N^{-\frac{2}{3}+\varepsilon}$. Here, $\kappa = |E - E_-|$ is given in (2.14).

Let γ_j be the j-th N-quantile of $\mu_\alpha \boxplus \mu_\beta$, *i.e.*, γ_j is the smallest real number such that

$$\mu_\alpha \boxplus \mu_\beta\big((-\infty, \gamma_j]\big) = \frac{j}{N}. \tag{2.20}$$

Similarly, we define γ_j^* to be the j-th N-quantile of $\mu_A \boxplus \mu_B$.

The following theorem is on the rigidity property of the eigenvalues of H.

Theorem 2.6 *(Rigidity at the lower edge). Suppose that Assumptions 2.1 and 2.2 hold. For any sufficiently small constant $c > 0$, we have that for all $1 \le i \le cN$,*

$$|\lambda_i - \gamma_i^*| \prec i^{-\frac{1}{3}} N^{-\frac{2}{3}}. \tag{2.21}$$

In fact, the same estimate also holds if γ_i^* is replaced with γ_i.

With the following additional assumptions on the upper edges of μ_α, μ_β and μ_A, μ_B, we can combine the current edge analysis with our strong local law in the bulk regime in [5]. This yields the rigidity result for the whole spectrum.

Assumption 2.7. We assume the following:
(ii') In a small δ-neighborhood of the upper edges of their supports, the measures μ_α and μ_β have a power law behavior, namely, there is a (large) constant $C \ge 1$ and exponents $-1 < t_+^\alpha, t_+^\beta < 1$ such that

$$C^{-1} \le \frac{\rho_\alpha(x)}{(E_+^\alpha - x)^{t_+^\alpha}} \le C\,, \qquad \forall x \in [E_+^\alpha - \delta, E_+^\alpha]\,,$$
$$C^{-1} \le \frac{\rho_\beta(x)}{(E_+^\beta - x)^{t_+^\beta}} \le C\,, \qquad \forall x \in [E_+^\beta - \delta, E_+^\beta]\,,$$

hold for some sufficiently small constant $\delta > 0$.
(v') For the upper edges of μ_A and μ_B, we have

$$\sup \operatorname{supp} \mu_A \le E_+^\alpha + \delta\,, \qquad \sup \operatorname{supp} \mu_B \le E_+^\beta + \delta\,,$$

for any constant $\delta > 0$ when N is sufficiently large.
(vii) The density function of $\mu_\alpha \boxplus \mu_\beta$ has a single interval support, *i.e.*,

$$\operatorname{supp} \mu_\alpha \boxplus \mu_\beta = [E_-, E_+]\,,$$

and has strictly positive density on (E_-, E_+).

Corollary 2.8 *(Rigidity for the whole spectrum). Suppose that Assumptions 2.1, 2.2 and 2.7 hold. Then we have, for all $i \in [\![1, N]\!]$, the estimate*

$$|\lambda_i - \gamma_i^*| \prec \max\left\{i^{-\frac{1}{3}}, (N-i+1)^{-\frac{1}{3}}\right\} N^{-\frac{2}{3}}. \tag{2.22}$$

The same estimate also holds if γ_i^* is replaced with γ_i. Moreover, we have the following estimate on the convergence rate of μ_H,

$$\sup_{x\in\mathbb{R}} \left|\mu_H((-\infty,x]) - \mu_A \boxplus \mu_B((-\infty,x])\right| \prec \frac{1}{N}. \tag{2.23}$$

The same estimate also holds if $\mu_A \boxplus \mu_B$ is replaced by $\mu_\alpha \boxplus \mu_\beta$.

Remark 2.9. In this work we focus on the extremal edges. Under Assumption 2.1 one can indeed prove [7] that $\mu_\alpha \boxplus \mu_\beta$ is supported on a single interval. In case that μ_α or μ_β are supported on several intervals, the free convolution $\mu_\alpha \boxplus \mu_\beta$ may also be supported on several intervals. In that case the presented work will directly apply to the smallest and largest edge points.

Remark 2.10. All of our results above hold also for the orthogonal setup, *i.e.*, when U is a random orthogonal matrix Haar distributed on the orthogonal group $\mathcal{O}(N)$. The proof is nearly the same as the unitary setup. A discussion on the necessary modification for the block additive model in the bulk regime can be found in Appendix C of [6]. Here for our model, the modification can be done in the same way. We omit the details.

3. Properties of the subordination functions at the regular edge

In this section, we collect some key properties of the subordination functions and related quantities, that will often be used in Sections 5-10. We first introduce

$$\begin{aligned}
\mathcal{S}_{AB} \equiv \mathcal{S}_{AB}(z) &:= (F_A'(\omega_B(z)) - 1)(F_B'(\omega_A(z)) - 1) - 1, \\
\mathcal{T}_A \equiv \mathcal{T}_A(z) &:= \frac{1}{2}\Big(F_A''(\omega_B(z))(F_B'(\omega_A(z)) - 1)^2 + F_B''(\omega_A(z))(F_A'(\omega_B(z)) - 1)\Big), \\
\mathcal{T}_B \equiv \mathcal{T}_B(z) &:= \frac{1}{2}\Big(F_B''(\omega_A(z))(F_A'(\omega_B(z)) - 1)^2 + F_A''(\omega_B(z))(F_B'(\omega_A(z)) - 1)\Big),
\end{aligned} \tag{3.1}$$

where we use the shorthand notation $F_A \equiv F_{\mu_A}$ and $F_B \equiv F_{\mu_B}$ for the negative reciprocal Stieltjes transforms of μ_A and μ_B, and where ω_A and ω_B are the subordination functions associated through (2.9). The main result in this section is the following proposition on the spectral domain $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$; see (2.15).

Proposition 3.1. *Suppose that Assumptions 2.1 and 2.2 hold. Then, for sufficiently small constant $\tau > 0$, we have the following statements:*

(*i*) *There exist strictly positive constants k and K, such that*

$$\min_i |a_i - \omega_B(z)| \geq k\,, \qquad \min_i |b_i - \omega_A(z)| \geq k\,, \tag{3.2}$$

$$\big|\omega_A(z)\big| \leq K, \qquad \big|\omega_B(z)\big| \leq K\,, \tag{3.3}$$

hold uniformly on $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$, for N sufficiently large.

(*ii*) *For the Stieltjes transform $m_{\mu_A \boxplus \mu_B}$ of $\mu_A \boxplus \mu_B$, we have that*

$$\operatorname{Im} m_{\mu_A \boxplus \mu_B}(z) \sim \begin{cases} \sqrt{\kappa+\eta}\,, & \text{if} \quad E \in \operatorname{supp} \mu_A \boxplus \mu_B\,, \\ \frac{\eta}{\sqrt{\kappa+\eta}}\,, & \text{if} \quad E \notin \operatorname{supp} \mu_A \boxplus \mu_B\,, \end{cases} \tag{3.4}$$

uniformly on $z = E + \mathrm{i}\eta \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$, for N sufficiently large, with κ given in (2.14).

(*iii*) *For $\mathcal{S}_{AB}$, $\mathcal{T}_A$ and $\mathcal{T}_B$ defined in (3.1), we have*

$$\mathcal{S}_{AB}(z) \sim \sqrt{\kappa+\eta}\,, \qquad |\mathcal{T}_A(z)| \leq C\,, \qquad |\mathcal{T}_B(z)| \leq C\,, \tag{3.5}$$

uniformly on $z \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$, for N sufficiently large, with some constant C. In addition, for $z = E + \mathrm{i}\eta \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$ with $|E - E_-| \leq \delta$ and $\eta \leq \delta$ for some sufficiently small constant $\delta > 0$, we also have

$$|\mathcal{T}_A(z)| \geq c\,, \qquad |\mathcal{T}_B(z)| \geq c\,, \tag{3.6}$$

for N sufficiently large, with some strictly positive constant $c = c(\delta)$.

(*iv*) *For ω_A, ω_B and $\mathcal{S}_{AB}$ we have*

$$|\omega_A'(z)| \leq C\frac{1}{\sqrt{\kappa+\eta}}\,, \qquad |\omega_B'(z)| \leq C\frac{1}{\sqrt{\kappa+\eta}}\,, \qquad |\mathcal{S}_{AB}'(z)| \leq C\frac{1}{\sqrt{\kappa+\eta}}\,, \tag{3.7}$$

any $z \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$, for N sufficiently large, with some constant C.

The proof of Proposition 3.1 is split into two steps. In the first step, carried out in Subsection 3.1, we derive the analogous statements for the N-independent measures μ_α and μ_β. This step requires only Assumption 2.1. In the second step, carried out in Subsection 3.2, we show that the statements carry over to the N-dependent measures μ_A and μ_B under Assumption 2.2, for N sufficiently large.

3.1. Free convolution measure $\mu_\alpha \boxplus \mu_\beta$

In this subsection, we derive some properties of the free additive convolution of μ_α and μ_β. We will always assume that μ_α and μ_β satisfy Assumption 2.1. From Assumption 2.1 (*iii*) we know that

$$\sup_{z\in\mathbb{C}^+} |m_{\mu_\alpha\boxplus\mu_\beta}(z)| \le C. \tag{3.8}$$

In addition, under Assumption 2.1, we see from Theorem 2.3 and Remark 2.4 in [8] that $\omega_\alpha(z)$, $\omega_\beta(z)$ and $m_{\mu_\alpha\boxplus\mu_\beta}(z)$ can be extended continuously to $\mathbb{C}^+\cup\mathbb{R}$. This together with (3.8) implies that $\mu_\alpha\boxplus\mu_\beta$ is absolutely continuous with a continuous and bounded density function.

Recall from Assumption 2.1 that $\operatorname{supp}\mu_\alpha = [E_-^\alpha, E_+^\alpha]$ and $\operatorname{supp}\mu_\beta = [E_-^\beta, E_+^\beta]$. We introduce the spectral domain $\mathcal{E}\subset\mathbb{C}$ by setting

$$\mathcal{E} := \{z\in\mathbb{C}^+\cup\mathbb{R} \,:\, E_-^\alpha + E_-^\beta - 1 \le \operatorname{Re} z \le E_+^\alpha + E_+^\beta + 1\,, 0\le \operatorname{Im} z\le \eta_{\mathrm{M}}\}\,, \tag{3.9}$$

where $\eta_{\mathrm{M}}>0$ is any constant. By Lemma 3.1 in [27], we have that $\operatorname{supp}\mu_\alpha\boxplus\mu_\beta\subset\mathcal{E}\cap\mathbb{R}$.

Lemma 3.2. *There exists a constant C such that*

$$\sup_{z\in\mathcal{E}}(|\omega_\alpha(z)| + |\omega_\beta(z)|) \le C\,. \tag{3.10}$$

Proof. Let $L>\max\{|E_+^\alpha + E_+^\beta + 1|, |E_-^\alpha + E_-^\beta - 1|\}$ and $M>10$ be large numbers to be chosen later. We will argue by contradiction. Assume first that there is $z\in\mathcal{E}$ such that

$$|\omega_\alpha(z)| > LM\,, \qquad\qquad |\omega_\beta(z)| > L\,. \tag{3.11}$$

Then we have from (2.9) that

$$\frac{1}{\omega_\alpha(z)+\omega_\beta(z)-z} = -\int_{\mathbb{R}}\frac{\mathrm{d}\mu_\alpha(x)}{x-\omega_\beta(z)} = \frac{1}{\omega_\beta(z)} + O((\omega_\beta(z))^{-2})\,, \tag{3.12}$$

$$\frac{1}{\omega_\alpha(z)+\omega_\beta(z)-z} = -\int_{\mathbb{R}}\frac{\mathrm{d}\mu_\beta(x)}{x-\omega_\alpha(z)} = \frac{1}{\omega_\alpha(z)} + O((\omega_\alpha(z))^{-2})\,, \tag{3.13}$$

as $L\to\infty$. Thus we get from (3.13), as $z\in\mathcal{E}$, that in the same limit

$$\frac{\omega_\beta(z)}{\omega_\alpha(z)} = O\left((\omega_\alpha(z)^{-1}\right)\,. \tag{3.14}$$

But then we have from (3.11) and (3.14) that

$$\frac{L}{|\omega_\alpha(z)|} \le \frac{|\omega_\beta(z)|}{|\omega_\alpha(z)|} \le C\frac{1}{|\omega_\alpha(z)|}\,, \tag{3.15}$$

hence for L sufficiently large, we get a contradiction.

Next, assume that there is $z\in\mathcal{E}$ such that

$$|\omega_\alpha(z)| > LM\,, \qquad\qquad |\omega_\beta(z)| \le L\,. \tag{3.16}$$

Then we conclude from (2.9) that

$$\frac{1}{|m_{\mu_\alpha}(\omega_\beta(z))|} = |\omega_\alpha(z) + \omega_\beta(z) - z| \geq \frac{LM}{2}\,, \tag{3.17}$$

for M sufficiently large, where we used that $z \in \mathcal{E}$. On the other hand, the Stieltjes transform $m_{\mu_\alpha}(z)$ does not have any zeros in $\mathcal{E}$ as the support of μ_α is connected. Thus there is a constant $c > 0$, depending on L, such that $|m_{\mu_\alpha}(z')| \geq c$, for all $z' \in \mathbb{C}^+$ with $|z'| \leq L$. Hence, for M sufficiently large, we get a contradiction from (3.17).

Finally, as both, (3.11) and (3.16), have been ruled out, we can conclude that

$$|\omega_\alpha(z)| \leq LM\,, \qquad\qquad |\omega_\beta(z)| \leq L\,, \tag{3.18}$$

for all $z \in \mathcal{E}$. This completes the proof of Lemma 3.2. $\square$

Recall from (2.12) that $E_- = \inf \operatorname{supp} \mu_\alpha \boxplus \mu_\beta$. Recall further that, for any spectral parameter z, $\kappa = \kappa(z)$ defined in (2.14) is the distance of $\operatorname{Re} z$ to the endpoints of $\operatorname{supp}(\mu_\alpha \boxplus \mu_\beta)$.

Lemma 3.3. *Let $u \in \mathbb{R}$ with $u \leq E_-$, then we have*

$$\operatorname{Re}\omega_\alpha(u) \leq E_-^\beta\,, \qquad\qquad \operatorname{Re}\omega_\beta(u) \leq E_-^\alpha\,. \tag{3.19}$$

Moreover, $\operatorname{Re}\omega_\alpha$ *and* $\operatorname{Re}\omega_\beta$ *are monotone increasing on* $(-\infty, E_-)$.

Proof. We argue by contradiction. Assume that there exists y' with $y' \leq E_-$ such that $\operatorname{Re}\omega_\alpha(y') > E_-^\beta$. Then either $\operatorname{Re}\omega_\alpha(y') \in (E_-^\beta, E_+^\beta)$ or $\operatorname{Re}\omega_\alpha(y') \geq E_+^\beta$. In the first case, using that the imaginary part of the identity $m_{\mu_\alpha\boxplus\mu_\beta}(z) = m_\alpha(\omega_\beta(z))$, we conclude that $\operatorname{Im} m_{\mu_\alpha\boxplus\mu_\beta}(y') > 0$, *i.e.*, the density of $\mu_\alpha \boxplus \mu_\beta$ at y' is strictly positive. This contradicts the definition of E_- (as the lowest endpoint $\operatorname{supp} \mu_\alpha \boxplus \mu_\beta$).

In the second case, $\operatorname{Re}\omega_\alpha(y') \geq E_+^\beta$, we have

$$\operatorname{Re} m_{\mu_\beta}(\omega_\alpha(y')) = \int_{E_-^\beta}^{E_+^\beta} \frac{(x - \operatorname{Re}\omega_\alpha(y'))\mathrm{d}\mu_\beta(x)}{|x - \omega_\alpha(y')|^2} < 0\,. \tag{3.20}$$

However, since $\operatorname{Re} m_{\mu_\beta}(\omega_\alpha(y')) = \operatorname{Re} m_{\mu_\alpha\boxplus\mu_\beta}(y')$, we get a contradiction as

$$\operatorname{Re} m_{\mu_\alpha\boxplus\mu_\beta}(y') = \int_y^\infty \frac{\mathrm{d}\mu_\alpha \boxplus \mu_\beta(x)}{x - y'} > 0\,, \tag{3.21}$$

by the definition of E_-.

From the above, we get $\operatorname{Re}\omega_\alpha(y') \leq E_-^\beta$. Repeating the argument for ω_β, we obtain (3.19).

Finally, that $\operatorname{Re}\omega_\alpha$ and $\operatorname{Re}\omega_\beta$ are increasing on $(-\infty, E_-)$ follows from the observation that $\operatorname{Re} m_{\mu_\alpha\boxplus\mu_\beta}$ is increasing on $(-\infty, E_-)$, the subordination property $m_{\mu_\alpha\boxplus\mu_\beta}(z) = m_{\mu_\beta}(\omega_\alpha(z))$ and (3.20). The same argument shows that $\operatorname{Re}\omega_\alpha$ is increasing on $(-\infty, E_-)$. This finishes the proof of Lemma 3.3. $\square$

We next show that we actually have $\operatorname{Re}\omega_\alpha(E_-) \leq E_-^\beta - k_0$ and $\operatorname{Re}\omega_\beta(E_-) \leq E_-^\alpha - k_0$, for some constant $k_0 > 0$. Our argument relies on the following computational lemma.

Lemma 3.4. *Let $\omega = \lambda + \mathrm{i}\nu$, with $\nu \geq 0$ and $|\omega| \leq \vartheta$, for some small $\vartheta > 0$. Let $-1 < t < 1$. Then,*

$$\int_0^\vartheta \frac{x^t \,\mathrm{d}x}{(x-\lambda)^2 + \nu^2} \sim \begin{cases} \frac{\lambda^t}{\nu}\,, & \text{if} \quad \lambda > \nu\,, \\ |\omega|^{t-1} \sim \lambda^{t-1}\,, & \text{if} \quad \lambda < -\nu\,, \\ \nu^{t-1}\,, & \text{if} \quad \nu > |\lambda|\,. \end{cases} \tag{3.22}$$

Proof. Follows from elementary estimations. $\square$

Recall from (2.6) that $F_\mu(w) = -1/m_\mu(w)$, $w \in \mathbb{C}^+$, denotes the negative reciprocal Stieltjes transform of any probability measure μ. As $F_\mu : \mathbb{C}^+ \to \mathbb{C}^+$ is analytic, and since μ is a probability measure, it admits the Nevanlinna representation

$$F_\mu(z) - z = \operatorname{Re} F_\mu(\mathrm{i}) + \int_{\mathbb{R}} \left(\frac{1}{x-z} - \frac{x}{1+x^2} \right) \mathrm{d}\widehat{\mu}(x)\,, \tag{3.23}$$

where $\widehat{\mu}$ is a Borel measure on $\mathbb{R}$. Assuming in addition that μ is compactly supported, a large z-expansion of both sides of (3.23) reveals that

$$\int_{\mathbb{R}} x \,\mathrm{d}\mu(x) = \operatorname{Re} F_\mu(\mathrm{i}) - \int_{\mathbb{R}} \frac{x}{1+x^2} \,\mathrm{d}\widehat{\mu}(x)\,, \tag{3.24}$$

and

$$\widehat{\mu}(\mathbb{R}) = \int_{\mathbb{R}} x^2 \,\mathrm{d}\mu(x) - \Big(\int_{\mathbb{R}} x \,\mathrm{d}\mu(x) \Big)^2 . \tag{3.25}$$

Lemma 3.5. *Let μ be a probability measure on $\mathbb{R}$ which is absolutely continuous with respect to Lebesgue measure, is of bounded support and satisfies $m_\mu(x) \neq 0$, for all $x \in \mathbb{R} \backslash \operatorname{supp}\mu$. Let $\widehat{\mu}$ be a Borel measure on $\mathbb{R}$ such that (3.23) holds, then we have that*

$$\operatorname{supp}\mu = \operatorname{supp}\widehat{\mu}\,. \tag{3.26}$$

Proof. The proof is almost identical to the proof of Lemma 3.2 in [7], we repeat it here for convenience of the reader. Outside the support of μ, the Stieltjes transform m_μ extends continuously to the real line and is real valued there. Taking the imaginary parts in (3.23) and using that $F_\mu(z) = -1/m_\mu(z)$, we get

$$\frac{\operatorname{Im} m_\mu(z)}{|m_\mu(z)|^2} - \operatorname{Im} z = \operatorname{Im} F_\mu(z) - z = \int_{\mathbb{R}} \frac{\operatorname{Im} z}{|y-z|^2} \mathrm{d}\widehat{\mu}(y)\,. \tag{3.27}$$

Since $m_\mu(x) \neq 0$, for all $x \in \mathbb{R}\backslash\operatorname{supp}\mu$, we can take the limit $\operatorname{Im} z \searrow 0$ in (3.27), and hence conclude by the Stieltjes inversion formula that $\widehat{\mu}$ is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}\backslash\operatorname{supp}\mu_\alpha$ with vanishing density function. We conclude that $\operatorname{supp}\widehat{\mu} \subseteq \operatorname{supp}\mu$.

To conclude that $\operatorname{supp}\mu \subseteq \operatorname{supp}\widehat{\mu}$ we argue by contradiction: Suppose that $\operatorname{supp}\widehat{\mu}$ is a proper subset of $\operatorname{supp}\mu$. Then there is a non-empty open interval $I \subset \operatorname{supp}\mu\backslash\operatorname{supp}\widehat{\mu}$ such that $f(\omega) := F_\mu(\omega) - \omega : \mathbb{C}^+ \to \mathbb{C}^+$ extends continuously to I with $\operatorname{Im} f(\omega) = 0$, for all $\omega \in I$. Hence by the Schwarz reflection principle, f extends analytically through I and m_μ is meromorphic on I. However, since $I \subset \operatorname{supp}\mu$, we have $\lim_{\eta\searrow 0} \operatorname{Im} m_\mu(\omega + \mathrm{i}\eta) > 0$ by Assumption 2.1, for almost all $\omega \in I$. Since m_μ is meromorphic on I and $\operatorname{Im} f(\omega) = \operatorname{Im} m_\mu(\omega)/|m_\mu(\omega)|^2$, $\omega \in I$, we hence also have $\lim_{\eta\searrow 0} \operatorname{Im} f(\omega + \mathrm{i}\eta) > 0$ for almost all $\omega \in I$, a contradiction to $\operatorname{Im} f(\omega) = 0$, for all $\omega \in I$. We conclude that I is empty and we have $\operatorname{supp}\widehat{\mu} = \operatorname{supp}\mu$. This proves (3.26). □

Remark 3.6. The assumptions of Lemma 3.5 are satisfied for μ_α and μ_β as follows easily from Assumption 2.1. Note that $m_{\mu_\alpha}(x) \neq 0$ for $x \in \mathbb{R}\backslash\operatorname{supp}\mu_\alpha$ is guaranteed by the single interval support condition. However, the condition that $m_{\mu_A}(x) \neq 0$, respectively $m_{\mu_B}(x) \neq 0$ cannot be guaranteed. But in this case we have the following inclusions for the supports of the N-dependent measures $\widehat{\mu}_A$ and $\widehat{\mu}_B$:

$$\operatorname{supp}\widehat{\mu}_A \subset I_{\mu_A}\,, \qquad \operatorname{supp}\widehat{\mu}_B \subset I_{\mu_B}\,, \tag{3.28}$$

where I_{μ_A}, I_{μ_B} is the smallest interval containing $\operatorname{supp}\mu_A$, $\operatorname{supp}\mu_B$. This easily follows from the proof of Lemma 3.5 by noticing that $m_{\mu_A}(x) \neq 0$, for $x \in \mathbb{R}\backslash I_{\mu_A}$, since $E \mapsto \operatorname{Re} m_{\mu_A}(E)$ is monotone on that domain, and similar for m_{μ_B}.

Lemma 3.7. *There is a constant $k_0 > 0$, such that*

$$\operatorname{Re}\omega_\alpha(E_-) \leq E_-^\beta - k_0\,, \qquad \operatorname{Re}\omega_\beta(E_-) \leq E_-^\alpha - k_0\,. \tag{3.29}$$

Moreover, there exists a constant C, such that

$$\operatorname{Im}\omega_\alpha(z) + \operatorname{Im}\omega_\beta(z) \leq \eta + C\operatorname{Im} m_{\mu_\alpha\boxplus\mu_\beta}(z)\,, \tag{3.30}$$

for all $z \in \mathcal{E}$. The constants k_0 and C only depend on μ_α and μ_β.

Proof. Let $z \in \mathcal{E}$. Taking the imaginary part in the subordination equations (2.9) we get

$$\frac{\operatorname{Im}\omega_\alpha(z) + \operatorname{Im}\omega_\beta(z) - \operatorname{Im} z}{|\omega_\alpha(z) + \omega_\beta(z) - z|^2} = \operatorname{Im} m_{\mu_\alpha \boxplus \mu_\beta}(z) .$$

Thus we obtain

$$\begin{aligned}\operatorname{Im}\omega_\alpha(z) + \operatorname{Im}\omega_\beta(z) &= \operatorname{Im} z + |\omega_\alpha(z) + \omega_\beta(z) - z|^2 \operatorname{Im} m_{\mu_\alpha \boxplus \mu_\beta}(z) \\ &\leq \eta + C \operatorname{Im} m_{\mu_\alpha \boxplus \mu_\beta}(z) ,\end{aligned}$$

where we used Lemma 3.2 to get the inequality. This proves (3.30).

We move on to prove the estimates in (3.29). Using

$$\operatorname{Im} m_{\mu_\alpha \boxplus \mu_\beta}(z) = \operatorname{Im}\omega_\alpha(z) \int_{\mathbb{R}} \frac{\mathrm{d}\mu_\beta(x)}{|x - \omega_\alpha(z)|^2} = \operatorname{Im}\omega_\beta(z) \int_{\mathbb{R}} \frac{\mathrm{d}\mu_\alpha(x)}{|x - \omega_\beta(z)|^2} , \tag{3.31}$$

and (2.9), we can write

$$\begin{aligned}&\frac{\operatorname{Im} m_{\mu_\alpha \boxplus \mu_\beta}(z)}{\operatorname{Im} z}\bigg(\Big(\int_{\mathbb{R}} \frac{\mathrm{d}\mu_\alpha(x)}{|x - \omega_\beta(z)|^2}\Big)^{-1} + \Big(\int_{\mathbb{R}} \frac{\mathrm{d}\mu_\beta(x)}{|x - \omega_\alpha(z)|^2}\Big)^{-1}\bigg) - 1 \\ &= \frac{\operatorname{Im} m_{\mu_\alpha \boxplus \mu_\beta}(z)}{\operatorname{Im} z} \frac{1}{|m_{\mu_\alpha \boxplus \mu_\beta}(z)|^2} ,\end{aligned}$$

for all $z \in \mathcal{E} \cap \mathbb{C}^+$. Since $\operatorname{Im} m_{\mu_\alpha \boxplus \mu_\beta}(z)/\operatorname{Im} z > 0$, for all $z \in \mathcal{E} \cap \mathbb{C}^+$, we obtain

$$\frac{\Big|\int_{\mathbb{R}} \frac{\mathrm{d}\mu_\alpha(x)}{x - \omega_\beta(z)}\Big|^2}{\int_{\mathbb{R}} \frac{\mathrm{d}\mu_\alpha(x)}{|x - \omega_\beta(z)|^2}} + \frac{\Big|\int_{\mathbb{R}} \frac{\mathrm{d}\mu_\beta(x)}{x - \omega_\alpha(z)}\Big|^2}{\int_{\mathbb{R}} \frac{\mathrm{d}\mu_\beta(x)}{|x - \omega_\alpha(z)|^2}} \geq 1 , \tag{3.32}$$

for all $z \in \mathcal{E} \cap \mathbb{C}^+$, where we used the subordination equations to express $m_{\mu_\alpha \boxplus \mu_\beta}(z)$. To condense the notation we introduce the quantities

$$R_\alpha(\omega) := \frac{\Big|\int_{\mathbb{R}} \frac{\mathrm{d}\mu_\alpha(x)}{x - \omega}\Big|^2}{\int_{\mathbb{R}} \frac{\mathrm{d}\mu_\alpha(x)}{|x - \omega|^2}} , \qquad R_\beta(\omega) := \frac{\Big|\int_{\mathbb{R}} \frac{\mathrm{d}\mu_\beta(x)}{x - \omega}\Big|^2}{\int_{\mathbb{R}} \frac{\mathrm{d}\mu_\beta(x)}{|x - \omega|^2}} , \qquad \omega \in \mathbb{C}^+ . \tag{3.33}$$

Fix some small $\vartheta > 0$. Recalling Lemma 3.4, we observe that there is a constant $c > 0$ (depending on ϑ) such that

$$\int_{E_-^\beta}^{E_-^\beta + \vartheta} \frac{\mathrm{d}\mu_\beta(x)}{|x - \omega|^2} \geq c \begin{cases} \frac{(\operatorname{Re}\omega - E_-^\beta)^{t_-^\beta}}{\operatorname{Im}\omega} , & \text{if} \quad \operatorname{Re}\omega - E_-^\beta \geq \operatorname{Im}\omega , \\ |\operatorname{Re}\omega - E_-^\beta|^{t_-^\beta - 1} , & \text{if} \quad \operatorname{Re}\omega - E_-^\beta \leq -\operatorname{Im}\omega , \\ (\operatorname{Im}\omega)^{t_-^\beta - 1} , & \text{if} \quad \operatorname{Im}\omega > |\operatorname{Re}\omega - E_-^\beta| , \end{cases} \tag{3.34}$$

for all ω with $|\omega - E_-^\beta| \le \vartheta$. (Since $-1 < t_-^\beta < 1$, the integral may be divergent in the limit $\operatorname{Im}\omega \to 0$, but this does not affect the following argument.)

Similarly, we have for $\omega \in \mathbb{C}$ satisfying $|\omega - E_-^\beta| \le \vartheta$,

$$\left| \int_{\mathbb{R}} \frac{\mathrm{d}\mu_\beta(x)}{x-\omega} \right| \le C + C' \int_0^\vartheta \frac{x^{t_-^\beta}}{|x + E_-^\beta - \omega|} \mathrm{d}x \le C + C' \left| E_-^\beta - \omega \right|^{t_-^\beta}, \tag{3.35}$$

for some strictly positive constants C and C' depending on ϑ. In particular, for $t_-^\beta \in [0,1)$, the right side of (3.35) is bounded. The inequalities (3.34) and (3.35) also hold true, upon possibly adjusting the constants, with the roles of α and β interchanged. We remark that we used similar estimates in the proof of Lemma 3.12 in [7].

Next, we introduce the quantities

$$d_\alpha(z) := \operatorname{dist}(\omega_\alpha(z), \operatorname{supp}\mu_\beta), \qquad d_\beta(z) := \operatorname{dist}(\omega_\beta(z), \operatorname{supp}\mu_\alpha). \tag{3.36}$$

We now claim that there are constants $k_0 > 0$ and $\varrho > 0$ such that $d_\alpha(z) \ge k_0$ and $d_\beta(z) \ge k_0$ for all $z \in \mathbb{C}^+ \cup \mathbb{R}$ with $|z - E_-| \le \varrho$. We proceed by distinguishing two cases: First assume that there is a z with $|z - E_-| \le \varrho$ such that

$$d_\alpha(z) \le \epsilon k, \qquad d_\beta(z) > k, \tag{3.37}$$

for some small constants $k > 0$ and $\epsilon > 0$ to be chosen below.

For $t_-^\beta \ge 0$, we obtain from (3.34) and (3.35) that for such z, we have

$$R_\beta(\omega_\alpha(z)) \le C \begin{cases} (\operatorname{Re}\omega_\alpha(z) - E_-^\beta)^{1-t_-^\beta}, & \text{if} \quad |\operatorname{Re}\omega_\alpha(z) - E_-^\beta| \ge \operatorname{Im}\omega_\alpha(z), \\ (\operatorname{Im}\omega_\alpha(z))^{1-t_-^\beta}, & \text{if} \quad |\operatorname{Re}\omega_\alpha(z) - E_-^\beta| < \operatorname{Im}\omega_\alpha(z). \end{cases} \tag{3.38}$$

Either way, we have $R_\beta(\omega_\alpha(z)) \le C(d_\alpha(z))^{1-t_-^\beta} \le C(\epsilon k)^{1-t_-^\beta}$, where we used that $t_-^\beta < 1$.

For $-1 < t_-^\beta < 0$, we obtain from (3.34) and (3.35) that for z with $|z - E_-| \le \varrho$ and (3.37) satisfied,

$$R_\beta(\omega_\alpha(z)) \le C \begin{cases} \operatorname{Im}\omega_\alpha(z) |\operatorname{Re}\omega_\alpha(z) - E_-^\beta|^{t_-^\beta}, & \text{if} \quad \operatorname{Re}\omega_\alpha(z) - E_-^\beta \ge \operatorname{Im}\omega_\alpha(z), \\ |\operatorname{Re}\omega_\alpha(z) - E_-|^{1+t_-^\beta}, & \text{if} \quad \operatorname{Re}\omega_\alpha(z) - E_-^\beta \le -\operatorname{Im}\omega_\alpha(z), \\ (\operatorname{Im}\omega_\alpha(z))^{1+t_-^\beta}, & \text{if} \quad |\operatorname{Re}\omega_\alpha(z) - E_-^\beta| < \operatorname{Im}\omega_\alpha(z). \end{cases} \tag{3.39}$$

In all three cases we find that $R_\beta(\omega_\alpha(z)) \le C(d_\alpha(z))^{1+t_\beta} \le C(\epsilon k)^{1+t_-^\beta}$, where we used $1 + t_-^\beta > 0$.

Since $d_\beta(z) > k$ and since we assumed that μ_α is not a single point mass, we have by the Cauchy-Schwarz inequality that

$$R_\alpha(\omega_\beta(z)) = \frac{\left|\int_{\mathbb{R}} \frac{\mathrm{d}\mu_\alpha(x)}{x-\omega_\beta(z)}\right|^2}{\int_{\mathbb{R}} \frac{\mathrm{d}\mu_\alpha(x)}{|x-\omega_\beta(z)|^2}} \le 1 - C_S(k,\varrho)\,, \tag{3.40}$$

for some strictly positive constant $C_S(k,\varrho) > 0$ depending on k and ϱ (and μ_α). Hence,

$$R_\alpha(\omega_\beta(z)) + R_\beta(\omega_\alpha(z)) \le 1 - C_S(k,\varrho) + C'(\epsilon k)^{1-|t_-^\beta|}\,, \tag{3.41}$$

with C' depending on ϱ. Thus for $\epsilon < (C_S(k,\varrho)/C')^{1/(1-|t_-^\beta|)}/k$ we get a contradiction with (3.32), for any $k > 0$. Thus there is no z with $|z - E_-| \le \varrho$ such that (3.37) can hold.

Assume thus that there is a z with $|z - E_-| \le \varrho$ such that

$$d_\alpha(z) \le \epsilon k\,, \qquad\qquad d_\beta(z) \le k\,, \tag{3.42}$$

for some sufficiently small $k > 0$ chosen below and with ϵ depending on k as above. Following the argumentation in (3.38) and (3.39) with the roles of α and β interchanged, we find

$$R_\alpha(\omega_\beta(z)) = \frac{\left|\int_{\mathbb{R}} \frac{\mathrm{d}\mu_\alpha(x)}{x-\omega_\beta(z)}\right|^2}{\int_{\mathbb{R}} \frac{\mathrm{d}\mu_\alpha(x)}{|x-\omega_\beta(z)|^2}} \le C k^{1-|t_-^\alpha|}\,, \tag{3.43}$$

while at the same time we have $R_\beta(\omega_\alpha(z)) \le C(\epsilon k)^{1-|t_-^\beta|}$ as we had above, with the constants depending on ϱ. Hence choosing $k > 0$ sufficiently small, we get a contradiction with (3.32), and we exclude (3.42).

We can therefore conclude that, for $\epsilon > 0$ and $k > 0$ sufficiently small, we have for all z with $|z - E_-| \le \varrho$, that

$$d_\alpha(z) > \epsilon k\,, \qquad\qquad d_\beta(z) > k\,. \tag{3.44}$$

Choosing $z = E_-$ this proves together with (3.30) and (3.19) the estimates in (3.29) with $k_0 := \epsilon k$. This concludes the proof of Lemma 3.7. $\square$

Lemma 3.8. *The lowest endpoint E_- of* $\operatorname{supp} \mu_\alpha \boxplus \mu_\beta$ *is the smallest real solution to the equation*

$$(F'_{\mu_\alpha}(\omega_\beta(z)) - 1)(F'_{\mu_\beta}(\omega_\alpha(z)) - 1) = 1\,, \qquad z \in \mathbb{R}\,. \tag{3.45}$$

Moreover, there are constants $\kappa_0 > 0$ and $\eta_0 > 0$ such that

$$\operatorname{Im} m_{\mu_\alpha \boxplus \mu_\beta}(z) \sim \operatorname{Im} \omega_\alpha(z) \sim \operatorname{Im} \omega_\beta(z) \sim \begin{cases} \sqrt{\kappa+\eta}\,, & \text{if } E \ge E_-\,, \\ \frac{\eta}{\sqrt{\kappa+\eta}} & \text{if } E < E_-\,, \end{cases} \tag{3.46}$$

uniformly for all $z = E + \mathrm{i}\eta \in \mathcal{E}_0$ *where*

$$\mathcal{E}_0 := \left\{ z \in \mathbb{C} \, : \, -\kappa_0 \le \operatorname{Re} z - E_- \le \kappa_0, 0 \le \operatorname{Im} z \le \eta_0 \right\}. \tag{3.47}$$

Proof of Lemma 3.8. From Lemma 3.7 we know that $\operatorname{Re}\omega_\alpha(E_-) \le E_-^\beta - k_0$ and $\operatorname{Re}\omega_\beta(E_-) \le E_-^\alpha - k_0$, $k_0 > 0$. From the subordination equations (2.9) and (3.23), we have that

$$F_{\mu_\alpha \boxplus \mu_\beta}(z) = F_{\mu_\alpha}(\omega_\beta(z)) = \operatorname{Re} F_{\mu_\alpha}(\mathrm{i}) + \omega_\beta(z) + \int_{\mathbb{R}} \left(\frac{1}{x - \omega_\beta(z)} - \frac{x}{1+x^2} \right) \mathrm{d}\widehat{\mu}_\alpha(x) , \tag{3.48}$$

for a Borel measure $\widehat{\mu}_\alpha$ on $\mathbb{R}$ with, according to Lemma 3.5, $\operatorname{supp}\widehat{\mu}_\alpha = \operatorname{supp}\mu_\alpha$. Arguing as in the proof of Lemma 3.5, we notice that $u \in \mathbb{R}$ is an edge of the measure $\mu_\alpha \boxplus \mu_\beta$, if $\operatorname{Im} m_{\mu_\alpha \boxplus \mu_\beta}(u) = 0$ and $m_{\mu_\alpha \boxplus \mu_\beta}$ fails to be analytic at $u \in \mathbb{R}$. Analyticity breaks down if either $F_{\mu_\alpha \boxplus \mu_\beta}(u) = 0$ or, according to (3.48), if $\omega_\beta(u) \in \operatorname{supp}\widehat{\mu}_\alpha = \operatorname{supp}\mu_\alpha$, or if ω_β fails to be analytic at u. For the lowest edge at $u = E_-$, we can exclude $F_{\mu_\alpha \boxplus \mu_\beta}(u) = 0$ by (3.8) and also $\omega(u) \in \operatorname{supp}\mu_\alpha$ as $\operatorname{Re}\omega_\alpha(E_-) \le E_-^\beta - k_0$, $k_0 > 0$. Thus $E_- \in \mathbb{R}$ is the smallest point where ω_β is not analytic.

We next claim that ω_β is not analytic at $u \in \mathbb{R}$ if $(F'_{\mu_\alpha}(\omega_\beta(u)) - 1)(F'_{\mu_\beta}(\omega_\alpha(u)) - 1) = 1$. We argue as follows. From (3.23) we know that there is a Borel measure $\widehat{\mu}_\beta$ such that

$$F_{\mu_\beta}(\omega) = \operatorname{Re} F_{\mu_\beta}(\mathrm{i}) + \omega + \int_{\mathbb{R}} \left(\frac{1}{x - \omega} - \frac{x}{1+x^2} \right) \mathrm{d}\widehat{\mu}_\beta(x) , \tag{3.49}$$

and F_{μ_β} is analytic in a disk of radius k_0 centered at $\omega = \omega_\beta(E_-)$ by (3.29). Here we also used that $\operatorname{supp}\widehat{\mu}_\beta = \operatorname{supp}\mu_\beta$ by Lemma 3.5. It follows that

$$F'_{\mu_\beta}(\omega) = 1 + \int_{\mathbb{R}} \frac{\mathrm{d}\widehat{\mu}_\beta(x)}{(x-\omega)^2} , \tag{3.50}$$

and in particular that $F'_{\mu_\beta}(\omega_\alpha(E_-)) > 1$, since $\omega_\alpha(E_-)$ is real valued as E_- is the lower endpoint of the support of $\mu_\alpha \boxplus \mu_\beta$ (recall (3.30)). By the analytic inverse function theorem, the functional inverse $F_{\mu_\beta}^{(-1)}$ of F_{μ_β} is analytic in a neighborhood of $F_{\mu_\beta}(\omega_\alpha(E_-))$. Thus the function

$$\widetilde{z}(\omega) := -F_{\mu_\alpha}(\omega) + \omega + F_{\mu_\beta}^{(-1)} \circ F_{\mu_\alpha}(\omega) \tag{3.51}$$

is well-defined and analytic in a complex neighborhood of $\omega_\alpha(E_-) \in \mathbb{R}$. It follows from (2.9) that $\omega_\beta(z)$ is a solution $\omega = \omega_\beta(z)$ to the equation $z = \widetilde{z}(\omega)$ (with $\operatorname{Im}\omega_\beta(z) \ge \operatorname{Im} z$). Moreover, we have $\omega_\alpha(z) = F_{\mu_\beta}^{(-1)} \circ F_{\mu_\alpha}(\omega_\beta(z))$.

The function $\widetilde{z}(\omega)$ admits the following Taylor expansion in a complex neighborhood of $\omega_\beta(E_-)$,

$$\widetilde{z}(\omega) = E_- + \widetilde{z}'(\omega_\beta(E_-))(\omega - \omega_\beta(E_-)) + \frac{1}{2}\widetilde{z}''(\omega_\beta(E_-))(\omega - \omega_\beta(E_-))^2 + O\left((\omega - \omega_\beta(E_-))^3\right) . \tag{3.52}$$

In particular, $\widetilde{z}(\omega)$ admits an inverse around $z = E_-$ that is locally analytic if and only if $\widetilde{z}'(\omega_\beta(E_-)) \neq 0$. Thus the smallest edge E_- of the support of $\mu_\alpha \boxplus \mu_\beta$, is the smallest $u \in \mathbb{R}$ such that $\widetilde{z}'(\omega_\beta(u)) = 0$. To find the location of the edge, we compute

$$\widetilde{z}'(\omega) = -F'_{\mu_\alpha}(\omega) + 1 + \frac{1}{F'_{\mu_\beta} \circ F^{(-1)}_{\mu_\beta} \circ F_{\mu_\alpha}(\omega)} F'_{\mu_\alpha}(\omega) . \tag{3.53}$$

Hence, choosing $\omega = \omega_\beta(z)$, we get

$$\widetilde{z}'(\omega_\beta(z)) = -F'_{\mu_\alpha}(\omega_\beta(z)) + 1 + \frac{1}{F'_{\mu_\beta}(\omega_\alpha(z))} F'_{\mu_\alpha}(\omega_\beta(z)) , \tag{3.54}$$

thence, from $\widetilde{z}'(\omega_\beta(E_-)) = 0$ we have

$$(F'_{\mu_\alpha}(\omega_\beta(E_-)) - 1)(F'_{\mu_\beta}(\omega_\alpha(E_-)) - 1) = 1 . \tag{3.55}$$

This proves (3.45).

We move on to proving (3.46). From (3.51) we compute,

$$\widetilde{z}''(\omega) = -F''_{\mu_\alpha}(\omega) + \frac{1}{F'_{\mu_\beta} \circ F^{(-1)}_{\mu_\beta} \circ F_{\mu_\alpha}(\omega)} F''_{\mu_\alpha}(\omega) - \frac{1}{(F'_{\mu_\beta} \circ F^{(-1)}_{\mu_\beta} \circ F_{\mu_\alpha}(\omega))^3} \left(F''_{\mu_\beta} \circ F^{(-1)}_{\mu_\beta} \circ F_{\mu_\alpha}(\omega)\right) \cdot (F'_{\mu_\alpha}(\omega))^2 ,$$

and thus by choosing $\omega = \omega_\beta(z)$, we get

$$\widetilde{z}''(\omega_\beta(z)) = -F''_{\mu_\alpha}(\omega_\beta(z)) + \frac{1}{F'_{\mu_\beta}(\omega_\alpha(z))} F''_{\mu_\alpha}(\omega_\beta(z)) - \frac{1}{(F'_{\mu_\beta}(\omega_\alpha(z)))^3} F''_{\mu_\beta}(\omega_\alpha(z)) \cdot (F'_{\mu_\alpha}(\omega_\beta(z)))^2 .$$

This we can rewrite as

$$\widetilde{z}''(\omega_\beta(z)) = \frac{F''_{\mu_\alpha}(\omega_\beta(z))}{F'_{\mu_\beta}(\omega_\alpha(z))} \big(1 - F'_{\mu_\beta}(\omega_\alpha(z))\big) - \frac{1}{(F'_{\mu_\beta}(\omega_\alpha(z)))^3} F''_{\mu_\beta}(\omega_\alpha(z)) \cdot \big(F'_{\mu_\alpha}(\omega_\beta(z))\big)^2 . \tag{3.56}$$

Thus choosing $z = E_-$ and recalling (3.54) and (3.55), we get

$$\widetilde{z}''(\omega_\beta(E_-)) = \frac{F''_{\mu_\alpha}(\omega_\beta(E_-))}{F'_{\mu_\beta}(\omega_\alpha(E_-))}\big(1 - F'_{\mu_\beta}(\omega_\alpha(E_-))\big) - \frac{F''_{\mu_\beta}(\omega_\alpha(E_-))}{F'_{\mu_\beta}(\omega_\alpha(E_-))}\big(F'_{\mu_\alpha}(\omega_\beta(E_-)) - 1\big)^2 . \tag{3.57}$$

From (3.50), we directly get

$$\begin{aligned} F'_{\mu_\beta}(\omega_\alpha(E_-)) &= 1 + \int_{\mathbb{R}} \frac{\mathrm{d}\widehat{\mu}_\beta(x)}{(x-\omega_\alpha(E_-))^2} > 1 , \\ F'_{\mu_\alpha}(\omega_\beta(E_-)) &= 1 + \int_{\mathbb{R}} \frac{\mathrm{d}\widehat{\mu}_\alpha(x)}{(x-\omega_\beta(E_-))^2} > 1 , \end{aligned} \tag{3.58}$$

where we used that $\widehat{\mu}_\alpha(\mathbb{R}) > 0$ and $\widehat{\mu}_\beta(\mathbb{R}) > 0$ as follows from (3.25) and the assumption that μ_α and μ_β are not single point masses. Moreover, recalling from (3.29) that $\omega_\alpha(E_-) \le E_-^\beta - k_0$, $\omega_\beta(E_-) \le E_-^\alpha - k_0$, we obtain

$$F''_{\mu_\beta}(\omega_\alpha(E_-)) = \int_{\mathbb{R}} \frac{\mathrm{d}\widehat{\mu}_\beta(x)}{(x-\omega_\alpha(E_-))^3} > 0 , \qquad F''_{\mu_\alpha}(\omega_\beta(E_-)) = \int_{\mathbb{R}} \frac{\mathrm{d}\widehat{\mu}_\alpha(x)}{(x-\omega_\beta(E_-))^3} > 0 . \tag{3.59}$$

Thus we infer from (3.57), (3.58) and (3.59) that there are constants $c > 0$ and $C < \infty$ such that

$$-C \le \widetilde{z}''(\omega_\beta(E_-)) \le -c . \tag{3.60}$$

Choosing $\omega = \omega_\beta(z)$ (thus $\widetilde{z}(\omega_\beta(z)) = z$) and using $\widetilde{z}'(\omega_\beta(E_-)) = 0$, $\widetilde{z}''(\omega_\beta(E_-)) < 0$ in (3.52), we get

$$\omega_\beta(z) - \omega_\beta(E_-) = \sqrt{\frac{-2}{\widetilde{z}''(\omega_\beta(E_-))}}\sqrt{E_- - z} + O(|z - E_-|) , \tag{3.61}$$

for z in a neighborhood of E_-. The branch of the square root is chosen such that $\operatorname{Im}\omega_\beta(z) > 0$, $z \in \mathbb{C}^+$.

Next, setting $z = E + \mathrm{i}\eta$, we observe that (3.60) and (3.61) imply, for z near E_-, that

$$\operatorname{Im}\omega_\beta(z) \sim \begin{cases} \sqrt{\kappa+\eta} , & \text{if } E \ge E_- , \\ \frac{\eta}{\sqrt{\kappa+\eta}} , & \text{if } E < E_- . \end{cases} \tag{3.62}$$

This proves the third estimate in (3.46). The second estimate is obtained in the same way by interchanging the roles of the indices α and β. Finally the first estimate follows from (3.31) and the fact that $\omega_\alpha(z)$ and $\omega_\beta(z)$, $z \in \mathcal{E}_0$, are away from the supports of the measure μ_β respectively μ_α by (3.29) and (3.61). This shows (3.46) and concludes the proof of Lemma 3.8. $\square$

Remark 3.9. From (3.61) and $m_{\mu_\alpha \boxplus \mu_\beta}(z) = m_{\mu_\alpha}(\omega_\beta(z))$ we get the precise behavior of $m_{\mu_\alpha \boxplus \mu_\beta}(z)$ on $\mathcal{E}_0$,

$$m_{\mu_\alpha \boxplus \mu_\beta}(z) - m_{\mu_\alpha \boxplus \mu_\beta}(E_-) = m'_{\mu_\alpha}(\omega_\beta(E_-))\sqrt{\frac{-2(\omega_\beta(E_-))}{\widetilde{z}''(\omega_\beta(E_-))}}\sqrt{E_- - z} + O(|z - E_-|)\,,$$

and thus by the Stieltjes inversion formula we have the square root behavior for the density of $\mu_\alpha \boxplus \mu_\beta$,

$$\mathrm{d}\mu_\alpha \boxplus \mu_\beta(x) \sim \sqrt{x - E_-}\,\mathrm{d}x\,, \qquad \forall x \in [E_-, E_- + \kappa_0]\,. \tag{3.63}$$

Corollary 3.10. *Let $\mathcal{E}_0$ be as in (3.47). Then the following behaviors hold uniformly for $z \in \mathcal{E}_0$,*

$$m'_{\mu_\alpha \boxplus \mu_\beta}(z) \sim \frac{1}{\sqrt{|z - E_-|}}\,, \qquad m''_{\mu_\alpha \boxplus \mu_\beta}(z) \sim \frac{1}{|z - E_-|^{3/2}}\,, \tag{3.64}$$

$$\omega'_\alpha(z) \sim \frac{1}{\sqrt{|z - E_-|}}\,, \qquad \omega''_\alpha(z) \sim \frac{1}{|z - E_-|^{3/2}}\,, \tag{3.65}$$

and

$$F'_{\mu_\alpha}(\omega_\beta(z)) \sim 1\,, \qquad F''_{\mu_\alpha}(\omega_\beta(z)) \sim 1\,, \qquad F'''_{\mu_\alpha}(\omega_\beta(z)) \sim 1\,. \tag{3.66}$$

The same estimates hold true when the roles of the subscripts α and β are interchanged.

Proof. Having established (3.46) for the behavior of ω_α and ω_β around the smallest edge E_-, the behaviors in (3.64) follow directly. Using the subordination equations (2.9), we note that $F'_{\mu_\alpha}(\omega_\beta(z))\omega'_\beta(z) = F'_{\mu_\beta}(\omega_\alpha(z))\omega'_\alpha(z) = -m'_{\mu_\alpha \boxplus \mu_\beta}(z)/(m_{\mu_\alpha \boxplus \mu_\beta}(z))^2$, which together with (3.64) imply (3.65). Finally, (3.66) follows directly from the analyticity of F_{μ_β} and F_{μ_α} in neighborhood of $\omega_\alpha(E_-)$, respectively $\omega_\beta(E_-)$. □

Let us define a second subdomain $\mathcal{E}_{\kappa_0}$ of $\mathcal{E}$ by setting

$$\mathcal{E}_{\kappa_0} := \{z \in \mathcal{E} \,:\, E_-^\alpha + E_-^\beta - 1 \le \operatorname{Re} z - E_- \le \kappa_0\,, 0 \le \operatorname{Im} z \le \eta_{\mathrm{M}}\} \tag{3.67}$$

with κ_0 and η_{M} as in (3.47). Note that $\mathcal{E}_0 \subset \mathcal{E}_{\kappa_0} \subset \mathcal{E}$. We further introduce the functions

$$\begin{aligned}
\mathcal{S}_{\alpha\beta} &\equiv \mathcal{S}_{\alpha\beta}(z) := (F'_{\mu_\alpha}(\omega_\beta(z)) - 1)(F'_{\mu_\beta}(\omega_\alpha(z)) - 1) - 1\,,\\
\mathcal{T}_\alpha &\equiv \mathcal{T}_\alpha(z) := \frac{1}{2}\big(F''_{\mu_\alpha}(\omega_\beta(z))\big(F'_{\mu_\beta}(\omega_\alpha(z)) - 1\big)^2 + F''_{\mu_\beta}(\omega_\alpha(z))\big(F'_{\mu_\alpha}(\omega_\beta(z)) - 1\big)\big)\,,\\
\mathcal{T}_\beta &\equiv \mathcal{T}_\beta(z) := \frac{1}{2}\big(F''_{\mu_\beta}(\omega_\alpha(z))\big(F'_{\mu_\alpha}(\omega_\beta(z)) - 1\big)^2 + F''_{\mu_\alpha}(\omega_\beta(z))\big(F'_{\mu_\beta}(\omega_\alpha(z)) - 1\big)\big)\,,\\
&\quad z \in \mathbb{C}^+\,.
\end{aligned} \tag{3.68}$$

These functions are essentially the first and second order derivatives of the subordination equations (2.9). We have the following corollary on the estimates of $m_{\mu_\alpha \boxplus \mu_\beta}$, ω_α, ω_β and also the above functions.

Corollary 3.11. *Let $\mathcal{E}_{\kappa_0}$ be as in (3.67) and let $\mathcal{E}_0$ be as in (3.47). Then*

$$\operatorname{Im} m_{\mu_\alpha \boxplus \mu_\beta}(z) \sim \operatorname{Im} \omega_\alpha(z) \sim \operatorname{Im} \omega_\beta(z) \sim \begin{cases} \sqrt{\kappa + \eta}\,, & \text{if } E \geq E_-\,, \\ \frac{\eta}{\sqrt{\kappa+\eta}}\,, & \text{if } E < E_-\,, \end{cases} \tag{3.69}$$

and

$$\mathcal{S}_{\alpha\beta}(z) \sim \sqrt{\kappa + \eta} \tag{3.70}$$

hold uniformly for $z \in \mathcal{E}_{\kappa_0}$, with κ given in (2.14). Moreover, we have

$$\mathcal{T}_\alpha(z) \sim 1\,, \qquad \mathcal{T}_\beta(z) \sim 1\,, \tag{3.71}$$

uniformly for $z \in \mathcal{E}_0$, respectively

$$|\mathcal{T}_\alpha(z)| \leq C\,, \qquad |\mathcal{T}_\beta(z)| \leq C\,, \tag{3.72}$$

uniformly for $z \in \mathcal{E}_{\kappa_0}$, for some constant C.

Proof of Corollary 3.11. Having established (3.46) for the behavior of ω_α and ω_β on $\mathcal{E}_0$, the behaviors in (3.69), (3.70) and (3.71) can be checked by elementary computations using Taylor expansions as in the proof of Lemma 3.8, and the estimates in (3.58) and (3.59).

Consider now the complementary domain $\mathcal{E}_{\kappa_0} \setminus \mathcal{E}_0$. Observe that $\kappa + \eta \sim 1$ in $\mathcal{E}_{\kappa_0} \setminus \mathcal{E}_0$. Hence, we have

$$\operatorname{Im} m_{\mu_\alpha \boxplus \mu_\beta}(z) = \int_{\mathbb{R}} \frac{\eta}{(x-E)^2 + \eta^2}\, \mathrm{d}\mu_\alpha \boxplus \mu_\beta(x) \sim \eta \tag{3.73}$$

uniformly on $\mathcal{E}_{\kappa_0} \setminus \mathcal{E}_0$. Then, from (3.30), (3.73) and $\operatorname{Im} \omega_\alpha(z) \geq \eta$, $\operatorname{Im} \omega_\beta(z) \geq \eta$, we get

$$\operatorname{Im} \omega_\alpha(z) \sim \eta\,, \qquad \operatorname{Im} \omega_\alpha(z) \sim \eta\,. \tag{3.74}$$

Observe that both estimates in (3.69) are of the same order as η if $z \in \mathcal{E}_{\kappa_0} \setminus \mathcal{E}_0$. Hence, we have (3.69).

Next, we show that (3.70) can be extended to the whole $\mathcal{E}_{\kappa_0} \setminus \mathcal{E}_0$. Since $\kappa + \eta \sim 1$, it suffices to show that the left side of (3.70) is comparable to 1 on $\mathcal{E}_{\kappa_0} \setminus \mathcal{E}_0$. We first consider real $z \in [E_-^\alpha + E_-^\beta - 1, E_-]$. Using (3.50) and its analogue for F'_{μ_α}, (3.55), (3.70), the monotonicity of $\omega_\alpha(z)$ and $\omega_\beta(z)$ on $(-\infty, E_- - \kappa_0]$ (*cf.*, Lemma 3.3), and (3.29), we see that

$$0 \le (F'_{\mu_\alpha}(\omega_\beta(z)) - 1)(F'_{\mu_\beta}(\omega_\alpha(z)) - 1) \le 1 - c\,, \qquad \forall z \in [E_-^\alpha + E_-^\beta - 1, E_- - \kappa_0]\,,$$

for some small constant $c > 0$. Hence, we have

$$\left|(F'_{\mu_\alpha}(\omega_\beta(z)) - 1)(F'_{\mu_\beta}(\omega_\alpha(z)) - 1) - 1\right| \sim 1\,, \qquad \forall z \in [E_-^\alpha + E_-^\beta - 1, E_- - \kappa_0]\,. \tag{3.75}$$

Then, (3.75) can be extended to all $z = E + \mathrm{i}\eta$, with $E \in [E_-^\alpha + E_-^\beta - 1, E_- - \kappa_0]$ and $0 \le \eta \le \widetilde{\eta}_0$ for sufficiently small constant $\widetilde{\eta}_0 > 0$ by continuity. This together with (3.70) gives the estimate in the regime $E \in [E_-^\alpha + E_-^\beta - 1, E_- + \kappa_0]$ and $0 \le \eta \le \eta_0$ after possibly reducing η_0 to $\widetilde{\eta}_0$ if $\eta_0 > \widetilde{\eta}_0$.

It remains to show that the left side of (3.70) is proportional to 1 when $E \in [E_-^\alpha + E_-^\beta - 1, E_- + \kappa_0]$ and $\eta_0 \le \eta \le \eta_{\mathrm{M}}$. To this end, we first recall (3.50), and observe from (3.48) that

$$\frac{\operatorname{Im} F_{\mu_\alpha}(\omega_\beta(z)) - \operatorname{Im}\omega_\beta(z)}{\operatorname{Im}\omega_\beta(z)} = \int_{\mathbb{R}} \frac{1}{|x - \omega_\beta(z)|^2}\,\mathrm{d}\widehat{\mu}_\alpha(x)\,. \tag{3.76}$$

Hence, using (3.50), (3.76) and their F_{μ_β} analogues, we have

$$\begin{aligned} &|(F'_{\mu_\alpha}(\omega_\beta(z)) - 1)(F'_{\mu_\beta}(\omega_\alpha(z)) - 1)| \\ &\quad \le \frac{\operatorname{Im} F_{\mu_\alpha}(\omega_\beta(z)) - \operatorname{Im}\omega_\beta(z)}{\operatorname{Im}\omega_\beta(z)}\,\frac{\operatorname{Im} F_{\mu_\beta}(\omega_\alpha(z)) - \operatorname{Im}\omega_\alpha(z)}{\operatorname{Im}\omega_\alpha(z)} \\ &\quad = \frac{\operatorname{Im}\omega_\beta(z) - \eta}{\operatorname{Im}\omega_\beta(z)}\,\frac{\operatorname{Im}\omega_\alpha(z) - \eta}{\operatorname{Im}\omega_\alpha(z)} \le 1 - c\,, \end{aligned} \tag{3.77}$$

for a strictly positive constant c, where in the second step we used the second equation in (2.9) and in the last step we used that $\eta \ge \eta_0$ and (3.74). Then, from (3.77) we get (3.70) in the whole of $\mathcal{E}_{\kappa_0}$.

Similarly, the upper bound in (3.72) follows from (3.74), (3.29), the monotonicity in Lemma 3.3, and the continuity of ω_α and ω_β. Omitting the details, we conclude the proof of Corollary 3.11. □

At this stage we have completed the first step in the proof of Proposition 3.1. In the next subsection, we carry out the second step where we translate results obtained so far for μ_α and μ_β to the measures μ_A and μ_B by giving the actual proof of Proposition 3.1.

3.2. Proof of Proposition 3.1

Consider the N-dependent measures μ_A and μ_B while always assuming that they satisfy Assumption 2.2. Let $\omega_A(z)$ and $\omega_B(z)$ denote the subordination functions associated by (2.11) to the measures μ_A and μ_B. Recall further the definition of the z-dependent quantities $\mathcal{S}_{AB}$, $\mathcal{T}_A$ and $\mathcal{T}_B$ in (3.1).

Recall that $E_- = \inf \operatorname{supp} \mu_\alpha \boxplus \mu_\beta$. Fix sufficiently small $\varepsilon, \delta > 0$ and let the domain $\mathcal{D}$ be defined by

$$\mathcal{D} := \mathcal{D}_{\text{in}} \cup \mathcal{D}_{\text{out}}\,, \tag{3.78}$$

with

$$\begin{aligned} \mathcal{D}_{\text{in}} &:= \{z \in \mathbb{C}^+ : |z - E_-| \le \delta\} \cap \{\operatorname{Im} z \ge N^{-1+10\varepsilon}, \operatorname{Re} z > E_- - N^{-1+10\varepsilon}\}\,, \\ \mathcal{D}_{\text{out}} &:= \{z \in \mathbb{C}^+ : |z - E_-| \le \delta\} \cap \{\operatorname{Re} z < E_- - N^{-1+10\varepsilon}\}\,. \end{aligned}$$

Notice that the bounds on A, B-quantities will be for spectral parameters z that are separated away from the limiting spectrum (*e.g.*, by assuming that $\operatorname{Im} z \ge N^{-1+10\varepsilon}$) unlike in case of the α, β-quantities.

Lemma 3.12. *Let μ_A, μ_B, μ_α and μ_β satisfy Assumptions 2.1 and 2.2. Then there is a constant C^* such that for any $z \in \mathcal{D}$ we have*

$$|\omega_A(z) - \omega_\alpha(z)| + |\omega_B(z) - \omega_\beta(z)| \le C^* \frac{N^{-1+\varepsilon}}{\sqrt{|z - E_-|}} \le N^{-1/2+\varepsilon}\,, \tag{3.79}$$

$$|\mathcal{S}_{AB}(z)| \sim \sqrt{|z - E_-|}\,, \tag{3.80}$$

and

$$|\mathcal{T}_A(z)| \sim 1\,, \qquad |\mathcal{T}_B(z)| \sim 1\,, \tag{3.81}$$

for N sufficiently large. Moreover, we have for any $z \in \mathcal{D}$ that

$$\operatorname{Im} m_{\mu_A \boxplus \mu_B}(z) \sim \sqrt{|z - E_-|}\,, \qquad z \in \mathcal{D}_{\text{in}}\,, \tag{3.82}$$

$$\operatorname{Im} m_{\mu_A \boxplus \mu_B}(z) \lesssim \frac{\operatorname{Im} z + O(N^{-1+\varepsilon})}{\sqrt{|z - E_-|}}\,, \qquad z \in \mathcal{D}_{\text{out}}\,, \tag{3.83}$$

for N sufficiently large. Furthermore, for the imaginary parts the bound (3.79) is, for N sufficiently large, sharpened to

$$|\operatorname{Im} \omega_A - \operatorname{Im} \omega_\alpha| + |\operatorname{Im} \omega_B - \operatorname{Im} \omega_\beta| \lesssim \frac{(\operatorname{Im} \omega_\alpha + \operatorname{Im} \omega_\beta) N^{-1+\varepsilon} + \operatorname{Im} z}{\sqrt{|z - E_-|}}\,, \tag{3.84}$$

for $z \in \mathcal{D}_{\rm out}$, $\eta \le N^{-1}$, *which implies that*

$$\inf \operatorname{supp} \mu_A \boxplus \mu_B \ge E_- - N^{-1+10\varepsilon} . \tag{3.85}$$

Away from the spectral edge we have the following weaker versions of (3.80), (3.81)*:*

$$|\mathcal{S}_{AB}(z)| \sim 1 , \tag{3.86}$$

$$|\mathcal{T}_A(z)| + |\mathcal{T}_B(z)| \le C , \tag{3.87}$$

hold uniformly for any z *with* $\delta \le |z - E_-| \le C$, *for* N *sufficiently large.*

Proof. We start by rewriting the subordination equation for μ_A and μ_B (*cf.*, (2.9) with $\mu_1 = \mu_A$, $\mu_2 = \mu_B$) as

$$\begin{aligned} F_{\mu_\alpha}(\omega_B(z)) - \omega_A(z) - \omega_B(z) + z &= r_1(z) , \\ F_{\mu_\beta}(\omega_A(z)) - \omega_A(z) - \omega_B(z) + z &= r_2(z) , \end{aligned} \tag{3.88}$$

where we introduced

$$r_1(z) := F_{\mu_\alpha}(\omega_B(z)) - F_{\mu_A}(\omega_B(z)) , \qquad r_2(z) := F_{\mu_\beta}(\omega_A(z)) - F_{\mu_B}(\omega_A(z)) . \tag{3.89}$$

We rely on the following local stability result for the system (3.88).

Lemma 3.13. *Let* $\omega_\alpha(z)$ *and* $\omega_\beta(z)$ *be the unique solutions to* (2.9) *with* $\mu_1 = \mu_\alpha$ *and* $\mu_2 = \mu_\beta$. *Fix* $z_0 \in \mathcal{D}$. *Assume that the functions* ω_A, $\omega_B : \mathbb{C}^+ \to \mathbb{C}^+$ *and* r_1, $r_2 : \mathbb{C}^+ \to \mathbb{C}$ *satisfy* (3.88) *with* $z = z_0$. *Assume moreover that there is a function* $q \equiv q(z)$ *such that*

$$|\omega_A(z_0) - \omega_\alpha(z_0)| \le q(z_0) , \qquad |\omega_B(z_0) - \omega_\beta(z_0)| \le q(z_0) , \tag{3.90}$$

with $q(z) = o(1)$ *and* $q(z)/\mathcal{S}_{\alpha\beta}(z) = o(1)$ *as* $N \to \infty$, *uniformly in* $z \in \mathcal{D}$, *where* $\mathcal{S}_{\alpha\beta}$ *is given in* (3.68). *Then we have*

$$|\omega_A(z_0) - \omega_\alpha(z_0)| + |\omega_B(z_0) - \omega_\beta(z_0)| \le C \frac{|r_1(z_0)| + |r_2(z_0)|}{|\mathcal{S}_{\alpha\beta}(z_0)|} , \tag{3.91}$$

for N *sufficiently large, with some constant* C *independent of* N *and* z_0.

Proof. The proof is similar to the proof of Proposition 4.1 in [3]. We start by Taylor expanding $F_{\mu_\alpha}(\omega_B(z_0))$ to second order around $\omega_\beta(z_0)$, so that the first equation in (3.88) reads

$$
\begin{aligned}
&F_{\mu_\alpha}(\omega_\beta(z_0)) + F'_{\mu_\alpha}(\omega_\beta(z_0))\big(\omega_B(z_0) - \omega_\beta(z_0)\big) - \omega_A(z_0) - \omega_B(z_0) + z_0 \\
&\quad = r_1(z_0) + O(|\omega_B(z_0) - \omega_\beta(z_0)|^2)\,,
\end{aligned}
$$

where we used that $F''_{\mu_\alpha}(\omega_\beta(z))$ is uniformly bounded for all $z \in \mathcal{D}$ (see (3.66)) and hence so is $F''_{\mu_\alpha}(\widetilde{\omega})$ for all $\widetilde{\omega} \in \mathbb{C}$ in a $q(z_0)$-neighborhood of $\omega_\beta(z_0)$. Subtracting from this last equation the subordination equation $F_{\mu_\alpha}(\omega_\beta(z_0)) - \omega_\alpha(z_0) - \omega_\beta(z_0) + z_0 = 0$, we obtain

$$
\big(F'_{\mu_\alpha}(\omega_\beta(z_0)) - 1\big)\Omega_B(z_0) - \Omega_A(z_0) = r_1(z_0) + O\left(|\Omega_B(z_0)|^2\right)\,, \tag{3.92}
$$

where we introduced $\Omega_B(z_0) := \omega_B(z_0) - \omega_\beta(z_0)$ and $\Omega_A(z_0) := \omega_A(z_0) - \omega_\alpha(z_0)$. Repeating the above with the roles of (α, A) and (β, B) interchanged, we also obtain

$$
\big(F'_{\mu_\beta}(\omega_\alpha(z_0)) - 1\big)\Omega_A(z_0) - \Omega_B(z_0) = r_2(z_0) + O\left(|\Omega_A(z_0)|^2\right)\,. \tag{3.93}
$$

Combining (3.92) and (3.93), we conclude that

$$
\begin{aligned}
|\Omega_A(z_0)| &\le C\frac{|r_1(z_0)| + |r_2(z_0)|}{|\mathcal{S}_{\alpha\beta}(z_0)|} + C'\frac{|\Omega_A(z_0)|^2}{|\mathcal{S}_{\alpha\beta}(z_0)|}\,, \\
|\Omega_B(z_0)| &\le C\frac{|r_1(z_0)| + |r_2(z_0)|}{|\mathcal{S}_{\alpha\beta}(z_0)|} + C'\frac{|\Omega_B(z_0)|^2}{|\mathcal{S}_{\alpha\beta}(z_0)|}\,,
\end{aligned} \tag{3.94}
$$

with $\mathcal{S}_{\alpha\beta}$ given in (3.68), for some numerical constant C and C' independent of N and z_0.

Next, since we are assuming that $|\Omega_A(z_0)|/|\mathcal{S}_{\alpha\beta}| \le q(z_0)/|\mathcal{S}_{\alpha\beta}| = o(1)$, and similarly for $\Omega_B(z_0)$, we conclude from (3.94) that

$$
|\Omega_A(z_0)| + |\Omega_B(z_0)| \le 4C\frac{|r_1(z_0)| + |r_2(z_0)|}{|\mathcal{S}_{\alpha\beta}(z_0)|}\,, \tag{3.95}
$$

which, upon renaming the constants was to be proved. □

Returning to the proof of Lemma 3.12, we use a continuity argument to prove (3.79) for spectral parameters $z \in \mathcal{D}$ close to the imaginary axis. First, for any $z \in \mathcal{D}$ with $\operatorname{Im} z = \eta_{\mathrm{M}}$, for some small fixed $\eta_{\mathrm{M}} \sim 1$, the local linear stability result of Lemma 4.2 of [3] shows that $|\omega_A(z) - \omega_\alpha(z)| + |\omega_B(z) - \omega_\beta(z)| \le 2|r_1(z)| + 2|r_2(z)| \le N^{-1+2\epsilon}$, provided that $\operatorname{Im}\omega_A(z) - \operatorname{Im} z \ge c > 0$ and $\operatorname{Im}\omega_B(z) - \operatorname{Im} z \ge c > 0$. These bounds follow from the subordination equation and the representation

$$
\operatorname{Im}\omega_A(z) - \operatorname{Im} z = \operatorname{Im} F_{\mu_A}(\omega_B(z)) - \operatorname{Im}\omega_B(z) = (\operatorname{Im} z)\int_{\mathbb{R}} \frac{\mathrm{d}\widehat{\mu}_A(x)}{|x - z|^2} \ge c' > 0\,, \tag{3.96}
$$

if $\operatorname{Im} z \ge \eta_{\mathrm{M}}$, and similarly for ω_B. Here we also used that $\widehat{\mu}_A(\mathbb{R}) > 0$; see (3.25).

Having established (3.79) for any $z \in \mathcal{D}$ with $\operatorname{Im} z = \eta_M$, we fix $E = \operatorname{Re} z$ and reduce the imaginary part of z. Let η_E be the smallest number such that (3.79) holds for any $z = E + \mathrm{i}\eta$ with $\eta \in [\eta_E, \eta_M]$; our goal is to show that $\eta_E \leq N^{-1+10\varepsilon}$. Suppose this is not the case. Use Lemma 3.13 with $q(z) := N^{-1+5\epsilon}/\sqrt{|z-E_-|}$ and with the choice $z_0 := E + \mathrm{i}(\eta_E - \zeta) \in \mathcal{D}$ for some tiny $\zeta > 0$. Since ω_A and ω_α are continuous (even analytic) at z_0, for a sufficiently small $\zeta > 0$, the estimate (3.79) for $z = E + \mathrm{i}\eta_E$ guarantees (3.90) for z_0. The other conditions of Lemma 3.13, namely $q(z) = o(1)$ and $q(z)/\mathcal{S}_{\alpha\beta}(z) = o(1)$, clearly hold since $\mathcal{S}_{\alpha\beta}(z) \sim \sqrt{|z-E_-|}$ by (3.70) and $|z - E_-| > N^{-1+10\epsilon}$ for $z \in \mathcal{D}$.

Hence (3.91) follows by Lemma 3.13. We will verify below that the right hand side of (3.91), with this choice of z_0, is smaller than the estimate $C^* N^{-1+\varepsilon}/\sqrt{|z_0 - E_-|}$ in (3.79). This shows that (3.79) holds for z_0 with imaginary part below η_E, contradicting to the definition of η_E. This proves (3.79) for all $z \in \mathcal{D}$.

It remains to bound the right side of (3.91) under (3.90). For $\delta > 0$ in (3.78) sufficiently small, it follows from (3.29) and (3.61), that $\operatorname{Re}\omega_\beta(z_0) < E_-^\alpha - k_0/2$, and hence under (3.90) that $\operatorname{Re}\omega_B(z_0) < E_-^\alpha - k_0/4$, for N sufficiently large. Let now I be a finite open interval such that $\operatorname{supp}\mu_A, \operatorname{supp}\mu_\alpha \subset I$, $\operatorname{dist}(\operatorname{Re}\omega_B(z_0), I) \geq k_0/8$. Let $h \in C^2(\mathbb{R})$, then integration by parts shows that

$$\Big|\int_{\mathbb{R}} h(x)\mathrm{d}\mu_A(x) - \int_{\mathbb{R}} h(x)\mathrm{d}\mu_\alpha(x)\Big| = \Big|\int_{\mathbb{R}} h'(x)(\mathcal{F}_{\mu_A}(x) - \mathcal{F}_{\mu_\alpha}(x))\mathrm{d}x\Big|, \tag{3.97}$$

where $\mathcal{F}_{\mu_A}$ and $\mathcal{F}_{\mu_\alpha}$ are the cumulative distribution functions of μ_A and μ_α. Thus, letting $s := d_L(\mu_A, \mu_\alpha)$, with d_L the Lévy distance, we get

$$\begin{aligned}\Big|\int_{\mathbb{R}} h(x)(\mathrm{d}\mu_A(x) - \mathrm{d}\mu_\alpha(x))\Big| &= \Big|\int_{\mathbb{R}} \big(h'(x) - h'(x+s)\big)\mathcal{F}_{\mu_A}(x)\mathrm{d}x \\ &\quad + \int_{\mathbb{R}} h'(x+s)\big(\mathcal{F}_{\mu_A}(x) - \mathcal{F}_{\mu_\alpha}(x+s)\big)\mathrm{d}x\Big| \\ &\leq Cs\Big(\sup_{x\in\mathbb{R}} |h''(x)| + \int_{\mathbb{R}} |h'(x)|\mathrm{d}x\Big), \end{aligned} \tag{3.98}$$

where we used that $|\mathcal{F}_{\mu_A}(x) - \mathcal{F}_{\mu_\alpha}(x+s)| \leq s$ from the definition of the Lévy distance. Let now χ be a smooth cut-off function such that $\chi(x) = 1$, if $x \in I$, and $\chi(x) = 0$, if $\operatorname{dist}(x, I) \geq \frac{k_0}{100}$. Then using (3.98) with $h(x) = \chi(x)(x - \omega_B(z_0))^{-1}$ we conclude from (3.98) that, under (3.90),

$$|m_{\mu_A}(\omega_B(z_0)) - m_{\mu_\alpha}(\omega_B(z_0))| \leq Cs \leq C\boldsymbol{d}, \tag{3.99}$$

where C depends only on k_0 and with $\boldsymbol{d}$ given in (2.3). Finally, by (3.29) we have

$$m_{\mu_\alpha}(\omega_\beta(E_-)) = \int_{\mathbb{R}} \frac{\mathrm{d}\mu_\alpha(x)}{x - \omega_\beta(E_-)} \geq c_1 > 0.$$

Furthermore, by the analyticity of the Stieltjes transform outside the support of μ_α and under (3.90) we conclude that $\operatorname{Re} m_{\mu_\alpha}(\omega_B(z_0)) \geq c_1/2$, for sufficiently small δ in (3.78). Similarly, $\operatorname{Re} m_{\mu_A}(\omega_B(z_0)) \geq c_1/3$ from (3.99) as $\boldsymbol{d} = o(1)$. Hence, recalling (3.89), we get under (3.90), that for sufficiently large N,

$$|r_1(z_0)| = |F_{\mu_A}(\omega_B(z)) - F_{\mu_\alpha}(\omega_B(z_0))| \leq \frac{|m_{\mu_A}(\omega_B(z_0)) - m_{\mu_\alpha}(\omega_B(z_0))|}{|m_{\mu_A}(\omega_B(z_0))m_{\mu_\alpha}(\omega_B(z_0))|} \leq C\boldsymbol{d} \leq CN^{-1+\varepsilon}\,. \tag{3.100}$$

Interchanging the roles of $(\mu_A, \mu_\alpha, \omega_B(z_0))$ and $(\mu_B, \mu_\beta, \omega_A(z_0))$, we obtain in the same way that $|r_2(z_0)| \leq C\boldsymbol{d} \leq CN^{-1+\varepsilon}$, assuming (3.90).

Finally, the denominator of (3.91) can be bounded as $\mathcal{S}_{\alpha\beta}(z_0) \sim \sqrt{|z_0 - E_-|}$ by (3.70). Thus we have

$$|\omega_A(z_0) - \omega_\alpha(z_0)| + |\omega_B(z_0) - \omega_\beta(z_0)| \lesssim \frac{\boldsymbol{d}}{|\mathcal{S}_{\alpha\beta}(z_0)|} \leq C^* \frac{N^{-1+\varepsilon}}{\sqrt{|z_0 - E_-|}} \lesssim N^{-1/2+\varepsilon}$$

if C^* is chosen sufficiently large. The last step uses that $|z_0 - E_-| > N^{-1+10\epsilon}$ since $z_0 \in \mathcal{D}$. This completes the proof of (3.79).

From this bound we can compare $\mathcal{S}_{\alpha\beta}$ and $\mathcal{S}_{AB}$, $\mathcal{T}_\alpha$ and $\mathcal{T}_A$, and $\mathcal{T}_\beta$ and $\mathcal{T}_B$, *e.g.*,

$$\begin{aligned}
&|\mathcal{S}_{AB}(z) - \mathcal{S}_{\alpha\beta}(z)| \\
&\quad\leq |(F'_{\mu_A}(\omega_B(z)) - 1)(F'_{\mu_B}(\omega_A(z)) - 1) - (F'_{\mu_A}(\omega_\beta(z)) - 1)(F'_{\mu_B}(\omega_\alpha(z)) - 1)| \\
&\qquad + |(F'_{\mu_A}(\omega_\beta(z)) - 1)(F'_{\mu_B}(\omega_\alpha(z)) - 1) - (F'_{\mu_\alpha}(\omega_\beta(z)) - 1)(F'_{\mu_\beta}(\omega_\alpha(z)) - 1)| \\
&\quad\lesssim |\omega_A(z) - \omega_\alpha(z)| + |\omega_B(z) - \omega_\beta(z)| + d \lesssim N^{-1/2+\varepsilon}\,, \qquad z \in \mathcal{D}\,,
\end{aligned}$$

(in the first estimate we used that F's are all regular and in the second we used the same in addition to (3.29) and (2.4)). Since $|\mathcal{S}_{\alpha\beta}| \geq N^{-1/2+5\varepsilon}$ in this regime, we immediately get (3.80). The bounds (3.81), (3.82), (3.83), (3.86) are proven exactly in the same way by showing that the difference between the finite-N quantity and the limiting quantity is smaller than the size of the limiting quantity given in (3.68) and (3.64).

The proof of (3.84) requires one more argument. Outside of the support, (3.79) is not optimal for the imaginary parts. Recall r_1 and r_2 from (3.89), $z \in \mathbb{C}^+$. Clearly

$$|\operatorname{Im} r_1(z)| \leq C(\operatorname{Im}\omega_B(z))N^{-1+\varepsilon}, \qquad |\operatorname{Im} r_2(z)| \leq C(\operatorname{Im}\omega_A(z))N^{-1+\varepsilon}\,, \qquad z \in \mathcal{D}\,,$$

since

$$\operatorname{Im} F_{\mu_\alpha}(\omega_B(z)) = \frac{\operatorname{Im} m_{\mu_\alpha}(\omega_B(z))}{|m_{\mu_\alpha}(\omega_B(z))|^2} = \frac{\operatorname{Im}\omega_B(z)}{|m_{\mu_\alpha}(\omega_B(z))|^2}\int_{\mathbb{R}} \frac{\mathrm{d}\mu_\alpha(x)}{|x - \omega_B(z)|^2}\,,$$

so changing α to A yields a factor $N^{-1+\varepsilon}$ by (2.3), since $\omega_B(z)$ is away from the support of μ_α as $\omega_\beta(z)$ is away from the support of μ_α and we have $|\omega_B(z) - \omega_\beta(z)| \leq N^{-1/2+\epsilon}$

by (3.79). Taking imaginary parts in (3.88) and using the representations from (3.23) yields the estimates

$$\operatorname{Im}\omega_B(z)\int_{\mathbb{R}}\frac{\mathrm{d}\widehat{\mu}_\alpha(x)}{|x-\omega_B(z)|^2}-\operatorname{Im}\omega_A(z)+\operatorname{Im}z=\operatorname{Im}r_1(z)=O\big(\operatorname{Im}\omega_B(z)N^{-1+\varepsilon}\big)\,,$$
$$\operatorname{Im}\omega_A(z)\int_{\mathbb{R}}\frac{\mathrm{d}\widehat{\mu}_\beta(x)}{|x-\omega_A(z)|^2}-\operatorname{Im}\omega_B(z)+\operatorname{Im}z=\operatorname{Im}r_2(z)=O\big(\operatorname{Im}\omega_A(z)N^{-1+\varepsilon}\big)\,,\tag{3.101}$$

$z\in\mathcal{D}$. Similarly, starting from the subordination equations for μ_α and μ_β, we have

$$\operatorname{Im}\omega_\beta(z)\int_{\mathbb{R}}\frac{\mathrm{d}\widehat{\mu}_\alpha(x)}{|x-\omega_\beta(z)|^2}-\operatorname{Im}\omega_\alpha(z)+\operatorname{Im}z=0\,,$$
$$\operatorname{Im}\omega_\alpha(z)\int_{\mathbb{R}}\frac{\mathrm{d}\widehat{\mu}_\beta(x)}{|x-\omega_\alpha(z)|^2}-\operatorname{Im}\omega_\beta(z)+\operatorname{Im}z=0\,.\tag{3.102}$$

In fact, using (3.79) we can change ω_B to ω_β and ω_A to ω_α in the integrands and error terms in (3.101), to get

$$\begin{aligned}&\operatorname{Im}\omega_B(z)\int_{\mathbb{R}}\frac{\mathrm{d}\widehat{\mu}_\alpha(x)}{|x-\omega_\beta(z)|^2}-\operatorname{Im}\omega_A(z)+\operatorname{Im}z\\&\quad=O\big(\operatorname{Im}\omega_\beta(z)N^{-1+\varepsilon}\big)+O\Big(\operatorname{Im}\omega_\beta(z)\frac{N^{-1+\epsilon}}{\sqrt{|z-E_-|}}\Big)\,,\\&\operatorname{Im}\omega_A(z)\int_{\mathbb{R}}\frac{\mathrm{d}\widehat{\mu}_\beta(x)}{|x-\omega_\alpha(z)|^2}-\operatorname{Im}\omega_B(z)+\operatorname{Im}z\\&\quad=O\big(\operatorname{Im}\omega_\alpha(z)N^{-1+\varepsilon}\big)+O\Big(\operatorname{Im}\omega_\alpha(z)\frac{N^{-1+\epsilon}}{\sqrt{|z-E_-|}}\Big)\,,\end{aligned}\tag{3.103}$$

$z\in\mathcal{D}$. Subtracting (3.102) from (3.103) and using that for very small η the determinant of the resulting linear system is very close to $\mathcal{S}_{\alpha\beta}(z)\sim\sqrt{|z-E_-|}$, $z\in\mathcal{D}$, from (3.80), we get

$$\begin{aligned}&|\operatorname{Im}\omega_A(z)-\operatorname{Im}\omega_\alpha(z)|+|\operatorname{Im}\omega_B(z)-\operatorname{Im}\omega_\beta(z)|\\&\quad\lesssim\frac{\operatorname{Im}\omega_\alpha(z)+\operatorname{Im}\omega_\beta(z)}{\sqrt{|z-E_-|}}N^{-1+\epsilon}+\frac{\operatorname{Im}\omega_\alpha(z)+\operatorname{Im}\omega_\beta(z)}{|z-E_-|}N^{-1+\epsilon}\,.\end{aligned}$$

Hence, recalling the bound in (3.69) for $\operatorname{Im}\omega_\alpha(z)$ and $\operatorname{Im}\omega_\beta(z)$, we get for $z\in\mathcal{D}_{\mathrm{out}}$ with $\eta\le N^{-1}$,

$$
\begin{aligned}
&|\operatorname{Im}\omega_A(z)-\operatorname{Im}\omega_\alpha(z)|+|\operatorname{Im}\omega_B(z)-\operatorname{Im}\omega_\beta(z)|\\
&\quad\lesssim \frac{\operatorname{Im}\omega_\alpha(z)+\operatorname{Im}\omega_\beta(z)}{\sqrt{|z-E_-|}}N^{-1+\epsilon}+\frac{\eta}{\sqrt{|z-E_-|}}\frac{N^{-1+\epsilon}}{|z-E_-|}\\
&\quad\lesssim \frac{\operatorname{Im}\omega_\alpha(z)+\operatorname{Im}\omega_\beta(z)}{\sqrt{|z-E_-|}}N^{-1+\epsilon}+\frac{\eta}{\sqrt{|z-E_-|}}\,.
\end{aligned}
$$

This proves (3.84).

Finally, to prove (3.85), let $z=x+\mathrm{i}\eta$ with $x\le E_- - N^{-1+10\varepsilon}$. At a distance of at least N^{-1} below E_-, we get

$$
\operatorname{Im} m_{\mu_\alpha\boxplus\mu_\beta}(z)=\operatorname{Im} z\int_{\mathbb{R}}\frac{\mathrm{d}\mu_\alpha\boxplus\mu_\beta(x)}{|x-z|^2}\le CN\operatorname{Im} z\,.
$$

Moreover from $m_{\mu_\alpha\boxplus\mu_\beta}(z)=m_\alpha(\omega_\beta(z))$ we have $\operatorname{Im} m_\alpha(\omega_\beta(z))\sim\operatorname{Im}\omega_\beta(z)$ since $\omega_\beta(z)$ is away from the support of μ_α. The same holds for $\omega_\alpha(z)$, so we get $\operatorname{Im}\omega_\alpha(z)+\operatorname{Im}\omega_\beta(z)\le CN\operatorname{Im} z$. Taking $\eta\searrow 0$, we note that the right hand side of (3.84) goes to zero. Thus we get $\operatorname{Im}\omega_A(x)=\operatorname{Im}\omega_B(x)=0$. Since $\operatorname{Im} m_{\mu_A\boxplus\mu_B}(z)\sim\operatorname{Im}\omega_A(z)$ in this regime, x cannot lie in the support of $\mu_A\boxplus\mu_B$. This proves (3.85). □

Recall that γ_j denoted the j-th N-quantile of $\mu_\alpha\boxplus\mu_\beta$ from (2.20) and similarly let γ_j^* denote the j-th N-quantile of $\mu_A\boxplus\mu_B$, *i.e.*, these are the smallest numbers γ_j and γ_j^* such that

$$
\mu_\alpha\boxplus\mu_\beta\big((-\infty,\gamma_j]\big)=\mu_A\boxplus\mu_B\big((-\infty,\gamma_j^*]\big)=\frac{j}{N}.
$$

Lemma 3.14 *(Rigidity). Suppose Assumptions 2.1 and 2.2 hold. Fix some sufficiently small $0<c<1$. Then we have, for any small $\epsilon>0$, the rigidity bound*

$$
|\gamma_j-\gamma_j^*|\le j^{-1/3}N^{-\frac{2}{3}+\varepsilon},\qquad j\in[\![1,cN]\!]\,, \tag{3.104}
$$

for N sufficiently large depending on ϵ and c.

Under the additional Assumption 2.7 we have the rigidity estimate for all quantiles, i.e.,

$$
|\gamma_j-\gamma_j^*|\le\min\{j^{-1/3},(N+1-j)^{-1/3}\}N^{-\frac{2}{3}+\varepsilon},\qquad j\in[\![1,N]\!]\,. \tag{3.105}
$$

Proof. The proof of these rigidity results are fairly straightforward from the information collected so far, by using standard arguments to translate the closeness of Stieltjes transform of two measures into closeness of their quantiles. We will just outline the argument. Recall the domain $\mathcal{E}_{\kappa_0}$ from (3.47).

First, we establish that there are at most N^ε γ_j-quantiles as well as N^ε γ_j^*-quantiles in an $N^{-2/3+\varepsilon}$ vicinity of $E_-=\inf\operatorname{supp}\mu_\alpha\boxplus\mu_\beta$. This fact is immediate for the γ_j

quantiles since their distribution is given by the regular square root law, see (3.63). For the γ_j^*-quantiles, we know from (3.85) that $\gamma_1^* \geq E_- - N^{-1+10\varepsilon}$. We compute from (3.82)

$$\begin{aligned}\frac{j}{N} &= \int_{-\infty}^{\gamma_j^*} \mathrm{d}\mu_A \boxplus \mu_B(x) = \int_{E_- - N^{-1+10\varepsilon}}^{\gamma_j^*} \mathrm{d}\mu_A \boxplus \mu_B(x) \\ &\leq C \int_{E_- - N^{-1+10\varepsilon}}^{\gamma_j^*} \operatorname{Im} m_{\mu_A \boxplus \mu_B}(x + \mathrm{i}N^{-1+10\varepsilon})\mathrm{d}x \\ &\leq C \int_{E_- - N^{-1+10\varepsilon}}^{\gamma_j^*} \big[|x - E_-| + N^{-1+10\varepsilon}\big]^{1/2}\mathrm{d}x \\ &\leq C|\gamma_j^* - E_-|^{3/2} + CN^{-1+10\varepsilon}|\gamma_j^* - E_-|,\end{aligned}$$

which means that

$$|\gamma_j^* - E_-| \geq c\Big(\frac{j}{N}\Big)^{2/3},$$

with some positive constant $c > 0$. So we have

$$\gamma_j^* \geq E_- + cN^{-2/3+\varepsilon}, \qquad \text{if} \quad j \geq cN^{3\varepsilon/2}, \tag{3.106}$$

and note that the condition $j \geq cN^{3\varepsilon/2}$ is equivalent to $\gamma_j \geq E_- + cN^{-2/3+\varepsilon}$. In the other direction we use

$$\int_{E_- - N^{-1+10\varepsilon}}^{\gamma_j^*} \mathrm{d}\mu_A \boxplus \mu_B(x) \geq c \int_{E_- - N^{-1+10\varepsilon}}^{\gamma_j^*} \operatorname{Im} m_{\mu_A \boxplus \mu_B}(x + \mathrm{i}N^{-1+10\varepsilon})\,\mathrm{d}x$$

if $|\gamma_j^* - E_-| \gg N^{-1+10\varepsilon}$. Using again (3.82) we get

$$\frac{j}{N} \geq c|\gamma_j^* - E_-|^{3/2}, \qquad i.e., \qquad \gamma_j^* \leq E_- + C\Big(\frac{j}{N}\Big)^{2/3} \qquad \forall j,$$

since this latter bound also holds in the case, when $|\gamma_j^* - E_-| \gg N^{-1+10\varepsilon}$ is not satisfied.

Thus we have established

$$|\gamma_j - \gamma_j^*| \leq |\gamma_j - E_-| + |\gamma_j^* - E_-| \leq CN^{-2/3+\varepsilon}, \qquad \text{whenever} \quad \gamma_j \leq E_- + N^{-2/3+\varepsilon}. \tag{3.107}$$

From the continuity of the free convolution (Proposition 4.13 of [10]) and the condition (2.3) we get

$$\mathrm{d}_L(\mu_A \boxplus \mu_B, \mu_\alpha \boxplus \mu_\beta) \leq \mathrm{d}_L(\mu_A, \mu_\alpha) + \mathrm{d}_L(\mu_B, \mu_\beta) \leq N^{-1+\epsilon}\,.$$

On the other hand, the definition of the Lévy distance and the boundedness of the density of $\mu_\alpha \boxplus \mu_\beta$ below $E_- + \kappa_0$ (see (3.63)) directly imply that

$$\big|\mu_A \boxplus \mu_B\big((-\infty, x)\big) - \mu_\alpha \boxplus \mu_\beta\big((-\infty, x)\big)\big| \leq CN^{-1+\varepsilon} \tag{3.108}$$

holds for any $x \leq E_- + \kappa_0$. Together with (3.107), this estimate immediately implies the bound (3.104).

For the proof of (3.105), we note that (ii') and (v') of Assumption 2.7 guarantee that near the upper edge of the support of $\mu_\alpha \boxplus \mu_\beta$ a similar rigidity statement holds as (3.104). Finally, (ii') of Assumption 2.7 together with the continuity and boundedness of the density of $\mu_\alpha \boxplus \mu_\beta$ (see (3.8)) imply that the density has a positive lower and upper bound away the two extreme edges of its support. This information together with (2.3) are sufficient to conclude that (3.108) hold uniformly for any $x \in \mathbb{R}$. The corresponding result (3.105) for the quantiles follows immediately. $\square$

Proof of Proposition 3.1. First, on the domain $\mathcal{D}$, (i) of Proposition 3.1 follows from (3.79), (3.29), the assumption (2.4) and also the continuity of ω_α and ω_β. In the complementary domain $\mathcal{D}_\tau(\eta_\mathrm{m}, \eta_\mathrm{M}) \setminus \mathcal{D}$, we first prove (3.3). Using the equations $m_{\mu_A \boxplus \mu_B} = m_{\mu_A}(\omega_B) = m_{\mu_B}(\omega_A)$, we see that the upper bounds on ω_A and ω_B follow from the fact that $|m_{\mu_A \boxplus \mu_B}(z)| \geq c$, which can easily be derived from the rigidity (3.104). For (3.2), we further split into two regimes. In the regime $\eta \geq \eta_0$ for some small $\eta > 0$, we use the fact $\mathrm{Im}\,\omega_A(z), \mathrm{Im}\,\omega_B(z) \geq \eta$ directly. In the regime $\eta \leq \eta_0$, we use the continuity of ω_A and ω_B, and also the monotonicity of the $\omega_A(u)$ and $\omega_B(u)$ for $u \in (-\infty, E_- - \delta]$ which can be proved similarly to the monotonicity of $\omega_\alpha(u)$ and $\omega_\beta(u)$ (*cf.*, (3.19)).

Similarly, on the domain $\mathcal{D}$, Proposition 3.1 (ii) follows from (3.82) and (3.83) directly. In the complementary domain $\mathcal{D}_\tau(\eta_\mathrm{m}, \eta_\mathrm{M}) \setminus \mathcal{D}$, we apply again the rigidity result (3.104) to conclude the proof.

Statement (iii) follows from (3.80), (3.81), (3.86) and (3.87).

Finally, to prove item (iv), we differentiate the subordination equations (2.9) with respect to z to get

$$\begin{pmatrix} 1 & 1 - F_A'(\omega_B(z)) \\ 1 - F_B'(\omega_A(z)) & 1 \end{pmatrix} \begin{pmatrix} \omega_A'(z) \\ \omega_B'(z) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

with the shorthand $F_A \equiv F_{\mu_A}$, $F_B \equiv F_{\mu_B}$. Hence,

$$\begin{pmatrix} \omega_A'(z) \\ \omega_B'(z) \end{pmatrix} = -\mathcal{S}^{-1}(z) \begin{pmatrix} F_A'(\omega_B(z)) \\ F_B'(\omega_A(z)) \end{pmatrix},$$

where $\mathcal{S} \equiv \mathcal{S}_{AB}$. Using (3.1) and (3.2) and (3.5), we directly get the first two estimates in (3.7), since $F_A'(\omega_B(z))$ and $F_B'(\omega_A(z))$ are uniformly bounded on $\mathcal{D}_\tau(\eta_\mathrm{m}, \eta_\mathrm{M})$ by (3.2).

Next, from the definition of $\mathcal{S}(z)$ in (3.1), we observe that

$$|\mathcal{S}'(z)| = \Big|F_B''(\omega_A)(F_A'(\omega_B)-1)\omega_A'(z) + F_A''(\omega_B)(F_B'(\omega_A)-1)\omega_B'(z)\Big| \leq C|\mathcal{S}^{-1}(z)|\,, \tag{3.109}$$

where in the second step we used (3.2) and the first two estimates in (3.7). Hence, by (3.5) we get the third estimate in (3.7) and statement (iv) is proved. This finishes the proof of Proposition 3.1. □

4. General structure of the proof

4.1. Partial randomness decomposition

In this subsection, we recall the partial randomness decomposition of the Haar unitary matrix used in [4], which will often be used below.

Let $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{iN})^T$ be the i-th column of the Haar distributed matrix U. Let θ_i be the argument of u_{ii}. The following partial randomness decomposition of U is taken from [16] (see also [24]): For any $i \in [\![1, N]\!]$, we can write

$$U = -\mathrm{e}^{\mathrm{i}\theta_i} R_i U^{\langle i\rangle}\,, \tag{4.1}$$

where $U^{\langle i\rangle}$ is a unitary block-diagonal matrix whose (i,i)-th entry equals 1, and its (i,i)-minor is Haar distributed on $\mathcal{U}(N-1)$. Hence, $U^{\langle i\rangle}\boldsymbol{e}_i = \boldsymbol{e}_i$ and $\boldsymbol{e}_i^* U^{\langle i\rangle} = \boldsymbol{e}_i^*$, where $\boldsymbol{e}_i$ is the i-th coordinator vector. Here R_i is a reflection matrix, defined as

$$R_i := I - \boldsymbol{r}_i\boldsymbol{r}_i^*\,, \tag{4.2}$$

where

$$\boldsymbol{r}_i := \sqrt{2}\frac{\boldsymbol{e}_i + \mathrm{e}^{-\mathrm{i}\theta_i}\boldsymbol{u}_i}{\|\boldsymbol{e}_i + \mathrm{e}^{-\mathrm{i}\theta_i}\boldsymbol{u}_i\|}\,. \tag{4.3}$$

Using $U^{\langle i\rangle}\boldsymbol{e}_i = \boldsymbol{e}_i$ and (4.1), we see that

$$\boldsymbol{u}_i = U\boldsymbol{e}_i = -\mathrm{e}^{\mathrm{i}\theta_i} R_i\boldsymbol{e}_i\,. \tag{4.4}$$

Hence, $R_i = R_i^*$ is actually the Householder reflection (up to a sign) sending $\boldsymbol{e}_i$ to $-\mathrm{e}^{-\mathrm{i}\theta_i}\boldsymbol{u}_i$. With the decomposition in (4.1), we can write

$$H = A + \widetilde{B} = A + R_i\widetilde{B}^{\langle i\rangle}R_i\,,$$

where we introduced the notations

$$\widetilde{B} := UBU^*, \qquad \widetilde{B}^{\langle i\rangle} := U^{\langle i\rangle}B(U^{\langle i\rangle})^*\,.$$

Observe that $\widetilde{B}^{\langle i\rangle}\boldsymbol{e}_i = b_i\boldsymbol{e}_i$ and $\boldsymbol{e}_i^*\widetilde{B}^{\langle i\rangle} = b_i\boldsymbol{e}_i^*$. Clearly, $\widetilde{B}^{\langle i\rangle}$ is independent of $\boldsymbol{u}_i$.

It is known that $\boldsymbol{u}_i \in S_{\mathbb{C}}^{N-1} := \{\boldsymbol{x} \in \mathbb{C}^N : \boldsymbol{x}^*\boldsymbol{x} = 1\}$ is a uniformly distributed complex vector, and there exists a Gaussian vector $\widetilde{\boldsymbol{g}}_i \sim \mathcal{N}_{\mathbb{C}}(0, N^{-1}I_N)$ such that

$$\boldsymbol{u}_i = \frac{\widetilde{\boldsymbol{g}}_i}{\|\widetilde{\boldsymbol{g}}_i\|}\,.$$

We then further introduce the notations

$$\boldsymbol{g}_i := \mathrm{e}^{-\mathrm{i}\theta_i}\widetilde{\boldsymbol{g}}_i\,, \qquad \boldsymbol{h}_i := \frac{\boldsymbol{g}_i}{\|\boldsymbol{g}_i\|} = \mathrm{e}^{-\mathrm{i}\theta_i}\boldsymbol{u}_i\,, \qquad \ell_i := \frac{\sqrt{2}}{\|\boldsymbol{e}_i + \boldsymbol{h}_i\|}\,. \tag{4.5}$$

Observe that the components g_{ik} of $\boldsymbol{g}_i$ are independent. Moreover, for $k \neq i$, $g_{ik} \sim N_{\mathbb{C}}(0, \frac{1}{N})$ while g_{ii} is a χ-distributed random variable with $\mathbb{E}g_{ii}^2 = \frac{1}{N}$. With the above notations, we can write $\boldsymbol{r}_i$ in (4.3) as

$$\boldsymbol{r}_i = \ell_i(\boldsymbol{e}_i + \boldsymbol{h}_i)\,. \tag{4.6}$$

In addition, using (4.4) and the fact $R_i^2 = I$, we have

$$R_i\boldsymbol{e}_i = -\boldsymbol{h}_i\,, \qquad R_i\boldsymbol{h}_i = -\boldsymbol{e}_i\,, \tag{4.7}$$

which also imply

$$\boldsymbol{h}_i^*\widetilde{B}^{\langle i\rangle}R_i = -\boldsymbol{e}_i^*\widetilde{B}\,, \qquad \boldsymbol{e}_i^*\widetilde{B}^{\langle i\rangle}R_i = -b_i\boldsymbol{h}_i^* = -\boldsymbol{h}_i^*\widetilde{B}\,. \tag{4.8}$$

Here, in the first equality of the second equation we used that $\boldsymbol{e}_i^*\widetilde{B}^{\langle i\rangle} = b_i\boldsymbol{e}_i$. We introduce the vectors

$$\mathring{\boldsymbol{g}}_i := \boldsymbol{g}_i - g_{ii}\boldsymbol{e}_i\,, \qquad \mathring{\boldsymbol{h}}_i := \frac{\mathring{\boldsymbol{g}}_i}{\|\boldsymbol{g}_i\|}\,,$$

where the χ-distributed variable g_{ii} is kicked out.

4.2. Summary of the proof route

In this subsection, we summarize the main route of the proof. While the final goal of the local law is to understand G_{ii}, $i \in [\![1, N]\!]$, and its averaged version, we work with several auxiliary quantities first. To understand their origin, it is useful to review the structure of our previous proofs of the local laws in the bulk [4,5]. We first introduce the following control parameters

$$\Psi \equiv \Psi(z) := \sqrt{\frac{1}{N\eta}}\,, \qquad \Pi \equiv \Pi(z) := \sqrt{\frac{\operatorname{Im} m_H}{N\eta}}\,, \tag{4.9}$$

for $z \in \mathbb{C}^+$, where $\eta = \operatorname{Im} z$. In [4], we investigated two main quantities:

$$S_i \equiv S_i(z) := \boldsymbol{h}_i^* \widetilde{B}^{\langle i \rangle} G \boldsymbol{e}_i \,, \qquad\qquad T_i \equiv T_i(z) := \boldsymbol{h}_i^* G \boldsymbol{e}_i \,. \tag{4.10}$$

In particular we showed that

$$S_i = \frac{z - \omega_B(z)}{a_i - \omega_B(z)} + O_\prec(\Psi) \,, \qquad\qquad T_i = O_\prec(\Psi) \,,$$

by performing integration by parts in the $\boldsymbol{h}_i^*$ variable. Using the identity

$$G_{ii} = \frac{1 - (\widetilde{B}G)_{ii}}{a_i - z}$$

and that

$$\begin{aligned}(\widetilde{B}G)_{ii} &= \boldsymbol{e}_i^* R_i \widetilde{B}^{\langle i \rangle} R_i G \boldsymbol{e}_i = -\boldsymbol{h}_i^* \widetilde{B}^{\langle i \rangle} R_i G \boldsymbol{e}_i = -S_i + \boldsymbol{h}_i^* \widetilde{B}^{\langle i \rangle} \boldsymbol{r}_i \boldsymbol{r}_i^* G \boldsymbol{e}_i \\ &= -S_i + \ell_i^2 (\boldsymbol{h}_i^* \widetilde{B}^{\langle i \rangle} \boldsymbol{h}_i + b_i h_{ii})(G_{ii} + T_i) \,,\end{aligned}$$

we obtained the entry-wise local law for G_{ii} from a precise control on S_i and T_i.

Technically S_i is a better quantity than G_{ii} to handle since integration by parts can be directly applied to it. However, along the calculation the quantity T_i appeared and a second integration by parts was needed to control it. We obtained a closed system of equations on the expectations of S_i and T_i (see (6.23)–(6.24) of [4]) from which the entry-wise local law in the bulk followed.

To obtain the law for the normalized trace of G in [5], we performed fluctuation averaging, but again not for G_{ii} directly. We considered averages (with arbitrary bounded weights d_i) of the quantity

$$Z_i := Q_i + G_{ii} \Upsilon \,,$$

where we defined

$$Q_i \equiv Q_i(z) := (\widetilde{B}G)_{ii} \operatorname{tr} G - G_{ii} \operatorname{tr} \widetilde{B}G \,, \tag{4.11}$$

$$\Upsilon \equiv \Upsilon(z) := \operatorname{tr} \widetilde{B}G - (\operatorname{tr} \widetilde{B}G)^2 + \operatorname{tr} G \operatorname{tr} \widetilde{B}G\widetilde{B} \,. \tag{4.12}$$

From the entry-wise laws it is clear that $|Q_i|, |\Upsilon| \prec \Psi$, and now we improve these bounds, at least in averaged sense in case of Q_i. Notice that Q_i is the most "symmetric" quantity, in particular $\sum_i Q_i = 0$, but technically it is not convenient to perform a high moment estimate for its weighted average, $\frac{1}{N}\sum_i d_i Q_i$. The reason is that one step of integration by parts generates an additional term, $G_{ii}\Upsilon$, which is hard to control directly. So instead of averaging Q_i, in [5] we included a counter term, *i.e.*, we averaged Z_i instead. We first proved that the average is one order better, *i.e.*,

$$\Big|\frac{1}{N}\sum_{i=1}^{N} d_i Z_i\Big| \prec \Psi^2 \tag{4.13}$$

with arbitrary deterministic weights $|d_i| \le 1$. Then, using (4.13) with $d_i \equiv 1$, we obtained $|\Upsilon| \prec \Psi^2$. Thus *a posteriori* we showed that the counter term $G_{ii}\Upsilon$ is irrelevant for estimates of order Ψ^2 and we obtained the same bound (4.13) for Q_i as well. Finally, the bounds on the average of Q_i with careful choices of the weights d_i and using the algebraic identities between G and $\widetilde{B}G$ yielded the averaged law for G_{ii} with the optimal $O_\prec(\Psi^2)$ error.

All results in [4,5] concerned the bulk. It is well known from the analogous results for Wigner matrices that the edge analysis is more difficult. The main reason is that the corresponding Dyson equation, the subordination equation in the current model, is unstable at the spectral edge, hence more precise estimates are necessary for the error terms. Theoretically, all error terms involving $\Psi = \frac{1}{\sqrt{N\eta}}$ should be improved by a factor of $\sqrt{\operatorname{Im} m}$, where we set $m := m_{\mu_A \boxplus \mu_B}$. This factor reflects that the density of states is small at the edge (at a square root edge we have $\operatorname{Im} m(z) \sim \sqrt{\kappa + \eta}$, where $\eta = \operatorname{Im} z$ and κ is the distance of $\operatorname{Re} z$ to the edge). This improvement exactly compensates for the bound of order $(\kappa+\eta)^{-1/2}$ on the inverse of the linearization of the subordination equation near the edge. However, this improvement is quite complicated to obtain and the method in [5] is not sufficient.

In this paper we present a new strategy to obtain the stronger bound. To prepare for the higher accuracy, already in the entry-wise law we work with two new quantities P_i and K_i instead of S_i and T_i. They are defined as

$$P_i \equiv P_i(z) := (\widetilde{B}G)_{ii}\operatorname{tr} G - G_{ii}\operatorname{tr}(\widetilde{B}G) + (G_{ii} + T_i)\Upsilon\,, \tag{4.14}$$

$$K_i \equiv K_i(z) := T_i + (b_i T_i + (\widetilde{B}G)_{ii})\operatorname{tr} G - (G_{ii} + T_i)\operatorname{tr}(\widetilde{B}G)\,. \tag{4.15}$$

We recognize that $P_i = Q_i + (G_{ii} + T_i)\Upsilon = Z_i + T_i\Upsilon$, *i.e.*, we included an additional counter term $T_i\Upsilon$ to the previous Z_i. While *a posteriori* this counter term turns out to be irrelevant, it is necessary in order to perform the integration by parts more precisely. Similarly,

$$K_i = \big(1 + b_i \operatorname{tr} G - \operatorname{tr}(\widetilde{B}G)\big)T_i + Q_i\,, \tag{4.16}$$

i.e., K_i is a linear combination of T_i and Q_i, it is nevertheless easier to work with K_i.

The proof of the estimates of the aforementioned quantities is divided into three parts. All these parts are performed for a fixed z and under the condition that G_{ii}'s and T_i's satisfy a weak a priori bound (*cf.*, (5.13)). This condition will be verified later in Section 8.

In the first part (Section 5) we obtain entry-wise bounds of the form

$$|K_i|, |Q_i|, |T_i|, |P_i| \prec \Psi, \qquad \text{as well as} \qquad |\Upsilon| \prec \Psi\,; \tag{4.17}$$

see Proposition 5.1. Notice that the estimates are still in terms of $\Psi = \frac{1}{\sqrt{N\eta}}$ without the improving factor $\sqrt{\operatorname{Im} m}$. These results would be possible to derive directly from the estimates in [4] by operating with S_i and T_i, we nevertheless use the new quantities, since the formulas derived along the entry-wise bounds will be used in the improved bounds later.

There is yet another reason for introducing the new quantities P_i and K_i, namely that in the current paper we have also changed the strategy concerning the entry-wise laws. In [4], a precursor to [5], we first proved entry-wise laws by deriving a system of equations for the expectation values (of S_i and T_i), complemented with concentration inequalities to enhance them to high probability bounds. For the improved bound on averaged quantities high moment estimates were performed only in [5], using the entry-wise law as an input. In the current paper we organize the proof in a more straightforward way, similarly to [6]. We bypass the fairly complicated concentration argument leading to the entry-wise law in [4] and we rely on high moment estimates directly even for the entry-wise law. This strategy is not only conceptually cleaner but also allows us to use essentially the same calculations for the entry-wise and the averaged law. The estimates of many error terms are shared in the two parts of the proofs; in case of some other estimates it will be sufficient to point out the necessary improvements. However, high moment estimates require to consider more carefully chosen quantities. For example, no direct high moment estimates are possible for S_i since it is even not a small quantity. But high moment estimates for the smaller quantities T_i and Q_i produce additional terms that are difficult to handle. It turns out that the carefully chosen counter terms in P_i and K_i make them suitable for performing high moment bounds.

More precisely, in the first step we compute the high moments of K_i and conclude that $|K_i| \prec \Psi$. In the second step we prove a high moment bound for $P_i = Q_i + (G_{ii} + T_i)\Upsilon$, *i.e.*, prove $|P_i| \prec \Psi$. In the third step we average this bound and conclude $|\Upsilon| \prec \Psi$, which in turn yields that $|Q_i| \prec \Psi$. Finally, from (4.16) we conclude that $|T_i| \prec \Psi$. This proves (4.17) and completes the entry-wise bounds.

In the second part of the proof (Section 6) we derive a rough bound on the averaged quantities. We will focus on $\frac{1}{N}\sum_i d_i Q_i$, since Q_i is the most fundamental quantity. Averaged quantities typically are one order better than the trivial entry-wise bounds indicate, *i.e.*, $|\frac{1}{N}\sum_i d_i Q_i| \prec \Psi^2 = (N\eta)^{-1}$, and indeed this was proven in [5] in the bulk and could be extended to the edge. In fact, due to the improvement at the edge, now we expect a bound of order $\Pi^2 \approx \operatorname{Im} m/N\eta$, but we cannot obtain this in general. In this second part of the proof, we therefore prove a bound of the form

$$\Big|\frac{1}{N}\sum_i d_i Q_i\Big| \prec \Pi\Psi \approx \frac{\sqrt{\operatorname{Im} m}}{N\eta},$$

which is “half-way” between the standard fluctuation averaging bound and the expected optimal bound. We compute the high moments of $\frac{1}{N}\sum_i d_i Q_i$ to achieve this bound. Interestingly, the apparently leading term in the high moment calculation already gives the optimal bound Π^2 (first term on the right of (6.5)), but a "cross-term" (when the derivative hits another factor of $\frac{1}{N}\sum_i d_i Q_i$) is responsible for the weaker $\Pi\Psi$ bound.

Another point to make is that it is not necessary to compute the high moments of another quantity for the rough averaged bound, unlike in [4,5] and in the first part of the current proof, where we always operated with two different quantities in parallel. Various error terms along the calculation of $\frac{1}{N}\sum_i d_i Q_i$ do contain T_i, but these terms can all be estimated using the entry-wise bound $|T_i| \prec \Psi$ only. Choosing a special weight sequence d_i we also improve the bound on Υ to $|\Upsilon| \prec \Pi\Psi$. In particular we could obtain an improved averaged bound on $P_i = Q_i + (G_{ii} + T_i)\Upsilon$ immediately, and with a little effort on K_i and T_i as well, but we do not need them.

Finally, in the third part of the proof (Section 7) we obtain the optimal Π^2 bound for the average of Q_i, but only for two very specially chosen weights, see (7.10)–(7.12). In fact, only the estimates on the "cross-term" need to be improved and the weights are carefully chosen to achieve an additional cancellation. Nevertheless, linear combinations of Q_i's with these two special sequences of weights are sufficient to imply an optimal self-consistent inequality for $\Lambda := |\Lambda_A| + |\Lambda_B|$ (see (7.2)).

The above three steps are done for a fixed z, under an a priori input on the bounds of G_{ii}'s and T_i's, (*cf.*, (5.13)). In order to get these inputs uniformly in $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$, we need to perform a continuity argument in the imaginary part of the spectral parameter $\eta = \operatorname{Im} z$. In Section 8 we will prove in Theorem 8.1 that

$$|P_{ii}| + |K_{ii}| + \Lambda_{\mathrm{d}}^c + \Lambda_T \prec \Psi\,, \qquad \text{and} \quad \Lambda_{\mathrm{d}} + \Lambda_A + \Lambda_B \prec \frac{1}{(N\eta)^{\frac{1}{3}}}\,, \tag{4.18}$$

uniformly in $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$. Note that the latter bound is weaker than our final goal of order $(N\eta)^{-1}$. Hence we call the second inequality in (4.18) *weak local law*, and the final bound (2.17), the optimal average law for G_{ii}, is called *strong local law* since it relies on the optimal $(N\eta)^{-1}$ bound in (4.18). The reason for this two-level approach, common in most local law proofs, is that the uniformity of the estimates in z is obtained by a continuity argument that cannot be optimally performed along the high moment estimates behind the fluctuation averaging. In fact, in the bulk regime, Theorem 2.6 of [4], (as well as for the analogous proofs for Wigner-type matrices) fluctuation averaging and high moment estimates were not even needed at this stage; a weak local law was obtained by a straightforward averaging of the entrywise law. The edge case is more subtle; Λ satisfies a quadratic inequality (see (8.3) later) which is linearly unstable. To counter this effect, we need stronger bounds on the error term. In the entrywise estimate for G_{ii}, our error bounds are given in terms of $\frac{1}{N\eta}\operatorname{Im}(G_{ii}+\mathcal{G}_{ii})$ (*cf.*, (5.3), (5.11), (5.56)) and we need to exploit the smallness of $\operatorname{Im}(G_{ii} + \mathcal{G}_{ii})$ via replacing it by its averaged (in i) version, $\operatorname{Im} m$. Since now our entrywise estimate itself is done via high moment bounds, pulling out the random and i-dependent error bound from the expectation and averaging it to get the improved bound for $\frac{1}{N}\sum_i d_i Q_i$ is not feasible. Hence, we need to perform a high moment estimate for $\frac{1}{N}\sum_i d_i Q_i$ independently to get the weak law

(Lemma 8.3). The point is that the weaker version of this high moment bound is still compatible with the continuity argument (Section 8), leading to the second estimate in (4.18). On the other hand (4.18) is enough to guarantee that the input (5.13) holds uniformly in $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$. With this uniform input, one can show that the discussion in the previous three steps hold also uniformly, leading eventually to the strong law uniformly in $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$.

We present the three parts explained above (Sections 5–7) first since they represent the essential and strongest ingredients of our proof. The proof of the weak law in Section 8 relies on similar steps, except that instead of *assuming* the controls on G_{ii} and T_i, they are *enforced* by inserting smooth cutoff functions. Thus, along the continuity argument we can bootstrap a single unconditional estimate (on the quantity with cutoffs), a procedure compatible with the high moment method. The cutoffs involve additional error terms that are still affordable as we are not aiming at the optimal bound for the moment.

At the end, in Section 9, by inverting the self-consistent inequality (7.2), we conclude that $\Lambda_\iota := \omega_\iota^c - \omega_\iota$, $\iota = A, B$, see (5.2) for the definition of ω_ι^c, are both stochastically dominated by Ψ^2. We finally notice that

$$\frac{1}{N}\sum_{i=1}^{N} d_i\Big(G_{ii} - \frac{1}{a_i - \omega_B^c}\Big)$$

may be expressed as a linear combination of the Q_i, see (9.5), this quantity is already stochastically bounded by $\Pi\Psi \leq \Psi^2$ from the second part of the proof. Since replacing ω_B^c with ω_B yields an error of at most Ψ^2, we obtain (2.17), the optimal average law for G_{ii}.

The actual proofs are considerably more complicated than this informal summary. On one hand, many error terms need to be estimated that have not been mentioned here, in particular we need fluctuation averaging with random weights, a novel complication that has not been considered before. On the other hand, in this summary we used the deterministic $\Psi = (N\eta)^{-1/2}$ and $\Pi \approx (\operatorname{Im} m/N\eta)^{1/2}$ as control parameters. In fact, Π is random, see (4.9), containing $\operatorname{Im} m_H$ which is $\operatorname{Im} m_{A\boxplus B}$ up to a random error that itself depends on Λ. Therefore an additional bootstrap for a fixed z is necessary to conclude a deterministic bound on Λ.

5. Entry-wise Green function subordination

In this section, we prove a subordination property for the Green function entries. From this section to Appendix C, without loss of generality, we assume that

$$\operatorname{tr} A = \operatorname{tr} B = 0\,. \tag{5.1}$$

We define the *approximate subordination functions* as

$$\omega_A^c(z) := z - \frac{\operatorname{tr} AG(z)}{m_H(z)}, \qquad \omega_B^c(z) := z - \frac{\operatorname{tr} \widetilde{B}G}{m_H(z)}, \qquad z \in \mathbb{C}^+. \tag{5.2}$$

It will be seen that the functions ω_A^c and ω_B^c are good approximations of ω_A and ω_B defined in (2.3) with $(\mu_1, \mu_2) = (\mu_A, \mu_B)$. Switching the roles of A and B, and also the roles of U and U^*, we introduce the following analogues of $\widetilde{B}$, H, and $G(z)$, respectively,

$$\widetilde{A} := U^* A U, \qquad \mathcal{H} := B + \widetilde{A}, \qquad \mathcal{G} \equiv \mathcal{G}(z) := (\mathcal{H} - z)^{-1}. \tag{5.3}$$

Observe that, by the cyclicity of the trace,

$$\omega_A^c(z) = z - \frac{\operatorname{tr} \widetilde{A}\mathcal{G}(z)}{m_H(z)}.$$

From (5.2) and the identity $(A + \widetilde{B} - z)G = I$, it is easy to check that

$$\omega_A^c(z) + \omega_B^c(z) - z = -\frac{1}{m_H(z)}, \qquad z \in \mathbb{C}^+. \tag{5.4}$$

Recall the quantities S_i and T_i defined in (4.10). We will also need their variants

$$\mathring{S}_i \equiv \mathring{S}_i(z) := \mathring{\boldsymbol{h}}_i^* \widetilde{B}^{\langle i \rangle} G \boldsymbol{e}_i = S_i - h_{ii} b_i G_{ii}, \qquad \mathring{T}_i \equiv \mathring{T}_i(z) := \mathring{\boldsymbol{h}}_i^* G \boldsymbol{e}_i = T_i - h_{ii} G_{ii}, \tag{5.5}$$

where the χ random variable h_{ii} is kicked out.

Further, we denote (dropping the z-dependence from the notation for brevity)

$$\Lambda_{\mathrm{d}i} := \left| G_{ii} - \frac{1}{a_i - \omega_B} \right|, \qquad \Lambda_{\mathrm{d}} := \max_i \Lambda_{\mathrm{d}i}, \qquad \Lambda_T := \max_i |T_i|. \tag{5.6}$$

We also define $\Lambda_{\mathrm{d}i}^c$ and Λ_{d}^c analogously by replacing ω_B by ω_B^c in the definitions of $\Lambda_{\mathrm{d}i}$ and Λ_{d}, respectively, *e.g.*,

$$\Lambda_{\mathrm{d}i}^c := \left| G_{ii} - \frac{1}{a_i - \omega_B^c} \right|, \qquad \Lambda_{\mathrm{d}}^c := \max_i \Lambda_{\mathrm{d}i}^c. \tag{5.7}$$

In addition, we use the notations $\widetilde{\Lambda}_{\mathrm{d}i}, \widetilde{\Lambda}_{\mathrm{d}}, \widetilde{\Lambda}_T, \widetilde{\Lambda}_{\mathrm{d}i}^c, \widetilde{\Lambda}_{\mathrm{d}}^c$ to represent their analogues, obtained by switching the roles of A and B, and the roles of U and U^*, in the definitions of $\Lambda_{\mathrm{d}i}, \Lambda_{\mathrm{d}}, \Lambda_T, \Lambda_{\mathrm{d}i}^c, \Lambda_{\mathrm{d}}^c$, *e.g.*,

$$\widetilde{\Lambda}_{\mathrm{d}i} := \left| \mathcal{G}_{ii} - \frac{1}{b_i - \omega_A} \right|, \qquad \widetilde{\Lambda}_{\mathrm{d}i}^c := \left| \mathcal{G}_{ii} - \frac{1}{b_i - \omega_A^c} \right|. \tag{5.8}$$

Recall P_i, K_i, and Υ defined in (4.14), (4.15) and (4.12). Note that all these quantities have analogues with tilde when the roles of A and B, and also the roles of U and U^* are switched.

We further observe the elementary identities

$$\widetilde{B}G = I - (A-z)G\,, \qquad\qquad G\widetilde{B} = I - G(A-z)\,. \tag{5.9}$$

Using the first identity in (5.9), we can rewrite Υ defined in (4.12) as

$$\Upsilon = \operatorname{tr} AG \operatorname{tr} \widetilde{B}G - \operatorname{tr} G \operatorname{tr} \widetilde{B}GA = \frac{1}{N}\sum_{i=1}^{N} a_i\Big(G_{ii}\operatorname{tr}\widetilde{B}G - (\widetilde{B}G)_{ii}\operatorname{tr} G\Big)\,. \tag{5.10}$$

To ease the presentation, we further introduce the control parameter

$$\Pi_i \equiv \Pi_i(z) := \sqrt{\frac{\operatorname{Im}(G_{ii}(z) + \mathcal{G}_{ii}(z))}{N\eta}}\,, \qquad\qquad i \in [\![1,N]\!]\,. \tag{5.11}$$

Note that since $\|H\| < \mathcal{K}$ (*cf.*, (2.13)), it is easy to see that $\operatorname{Im} G_{ii}(z) \gtrsim \eta$ and $\operatorname{Im} \mathcal{G}_{ii}(z) \gtrsim \eta$ for all $z \in \mathcal{D}_\tau(0,\eta_{\mathrm{M}})$, by spectral decomposition. This implies

$$\frac{1}{\sqrt{N}} \lesssim \Pi_i(z)\,, \qquad\qquad \forall z \in \mathcal{D}_\tau(0,\eta_{\mathrm{M}})\,. \tag{5.12}$$

In this section, we derive the following Green function subordination property. Recall the definitions of P_i and K_i in (4.14) and (4.15), as well as the definition of the control parameter Ψ in (4.9).

Proposition 5.1. *Suppose that the assumptions of Theorem 2.5 hold. Fix $z \in \mathcal{D}_\tau(\eta_{\mathrm{m}},\eta_{\mathrm{M}})$. Assume that*

$$\Lambda_{\mathrm{d}}(z) \prec N^{-\frac{\gamma}{4}}\,, \qquad \widetilde{\Lambda}_{\mathrm{d}}(z) \prec N^{-\frac{\gamma}{4}}\,, \qquad \Lambda_T(z) \prec 1\,, \qquad \widetilde{\Lambda}_T(z) \prec 1\,. \tag{5.13}$$

Then we have, for all $i \in [\![1,N]\!]$, that

$$|P_i(z)| \prec \Psi(z)\,, \qquad\qquad |K_i(z)| \prec \Psi(z)\,. \tag{5.14}$$

In addition, we also have that

$$|\Upsilon(z)| \prec \Psi(z) \tag{5.15}$$

and, for all $i \in [\![1,N]\!]$, that

$$\Lambda^c_{\mathrm{d}i}(z) \prec \Psi(z), \qquad\qquad |T_i| \prec \Psi(z)\,. \tag{5.16}$$

The same statements hold if we switch the roles of A and B, and also the roles of U and U^*.

Before the actual proof of Proposition 5.1, we establish several bounds that follow from the assumption in (5.13). From the definitions in (5.6), the assumptions in (5.13), together with (3.2), we see that

$$\max_{i\in[\![1,N]\!]}|G_{ii}|\prec 1\,, \qquad \max_{i\in[\![1,N]\!]}|T_i|\prec 1\,. \tag{5.17}$$

Analogously, we also have $\max_{i\in[\![1,N]\!]}|\mathcal{G}_{ii}|\prec 1$. Hence, under (5.13), we see that

$$\max_{i\in[\![1,N]\!]}\Pi_i(z)\prec\Psi(z).$$

Moreover, using the identities in (5.9), we also get from the first bound in (5.17) that

$$\max_{i\in[\![1,N]\!]}|(XGY)_{ii}|\prec 1, \qquad X,Y=I \text{ or } \widetilde{B}. \tag{5.18}$$

In addition, from (2.11) we see that

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{a_i-\omega_B(z)}=m_{\mu_A}(\omega_B(z))=m_{\mu_A\boxplus\mu_B}(z). \tag{5.19}$$

Then, the first bound in (5.13), together with (5.19), (5.9), (3.3) and (3.2), leads to the following estimates

$$\begin{aligned}\operatorname{tr}G&=m_{\mu_A\boxplus\mu_B}+O_\prec(N^{-\frac{\gamma}{4}})\,,\\ \operatorname{tr}\widetilde{B}G&=(z-\omega_B)m_{\mu_A\boxplus\mu_B}+O_\prec(N^{-\frac{\gamma}{4}})\,,\\ \operatorname{tr}\widetilde{B}G\widetilde{B}&=(\omega_B-z)\big(1+(\omega_B-z)m_{\mu_A\boxplus\mu_B}\big)+O_\prec(N^{-\frac{\gamma}{4}})\,.\end{aligned} \tag{5.20}$$

Furthermore, by (3.2), (3.3), and (5.19), we see that all the above tracial quantities are $O_\prec(1)$. This also implies that $|\Upsilon|\prec 1$, (*cf.*, (4.12)). Moreover, from (5.2) and the first two equations in (5.20), we can get the following rough estimate under (5.13) and (3.2),

$$\omega_B^c=\omega_B+O_\prec(N^{-\frac{\gamma}{4}})\,. \tag{5.21}$$

Further, we make the following convention in the rest of the paper: the notation $O_\prec(\Psi^k)$, for any given integer k, represents some generic (possibly) z-dependent random variable $X\equiv X(z)$ which satisfies

$$|X|\prec\Psi^k, \qquad \text{and} \qquad \mathbb{E}|X|^q\prec\Psi^{qk}\,,$$

for any given positive integer q. The first bound above follows from the original definition of the notation $O_\prec(\cdot)$ directly. It turns out that it is more convenient to require the second one in our discussions below as well. It will be clear that the second bound always follows from the first one whenever this notation will be used. For more details, we refer to the paragraph above Proposition 6.1 in [5]. Analogously, for all notation of the form $O_{\prec}(\Gamma)$ with some deterministic control parameter Γ, we make the same convention.

Proof of Proposition 5.1. To prove (5.14), it suffices to show the high order moment estimates

$$\mathbb{E}\left[|P_i|^{2p}\right] \prec \Psi^{2p}, \qquad \mathbb{E}\left[|K_i|^{2p}\right] \prec \Psi^{2p}, \tag{5.22}$$

for any fixed $p \in \mathbb{N}$. Let us introduce the notations

$$\mathfrak{m}_i^{(k,l)} := P_i^k \overline{P_i^l}, \qquad \mathfrak{n}_i^{(k,l)} := K_i^k \overline{K_i^l}, \qquad k,l \in \mathbb{N}, \qquad i \in [\![1,N]\!]. \tag{5.23}$$

With the definitions in (5.23) and the convention made after (5.21), we have the following recursive moment estimates. This type of estimates were used first in [23] to derive local laws for sparse Wigner matrices.

Lemma 5.2 *(Recursive moment estimate for P_i and K_i). Suppose the assumptions of Proposition 5.1. Then, for any fixed integer $p \geq 1$ and any $i \in [\![1,N]\!]$, we have*

$$\mathbb{E}[\mathfrak{m}_i^{(p,p)}] = \mathbb{E}[O_{\prec}(\Psi)\mathfrak{m}_i^{(p-1,p)}] + \mathbb{E}[O_{\prec}(\Psi^2)\mathfrak{m}_i^{(p-2,p)}] + \mathbb{E}[O_{\prec}(\Psi^2)\mathfrak{m}_i^{(p-1,p-1)}], \tag{5.24}$$

$$\mathbb{E}[\mathfrak{n}_i^{(p,p)}] = \mathbb{E}[O_{\prec}(\Psi)\mathfrak{n}_i^{(p-1,p)}] + \mathbb{E}[O_{\prec}(\Psi^2)\mathfrak{n}_i^{(p-2,p)}] + \mathbb{E}[O_{\prec}(\Psi^2)\mathfrak{n}_i^{(p-1,p-1)}], \tag{5.25}$$

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 $p = 1$.

Although in the statements of Lemma 5.2, we use Ψ, in the proof, we actually get better estimates in terms of Π_i^2 instead of Ψ^2 for some error terms. We will keep the stronger form of these estimates since the same errors will appear in the averaged bounds in Section 6 as well. The average of these errors is typically smaller than Ψ^2.

Proof of Lemma 5.2 . The proof is very similar to that of Lemma 7.3 of [6], which is presented for the block additive model in the bulk regime. It suffices to go through the strategy in [6] for our additive model again. The strategy also works well at the regular edge, provided (3.2) and (3.3) hold. In addition, instead of the control parameter Ψ used in the proof of Lemma 7.3 of [6], we aim here at controlling many errors in terms of Π_i. This requires a more careful estimate on the error terms. Due to the similarity to the proof of Lemma 7.3 of [6], we only sketch the proof of Lemma 5.2 in the sequel.

For each $i \in [\![1,N]\!]$, we write

$$\mathbb{E}[\mathfrak{m}_i^{(p,p)}] = \mathbb{E}[P_i\mathfrak{m}_i^{(p-1,p)}] = \mathbb{E}[(\widetilde{B}G)_{ii}\mathrm{tr}\, G\mathfrak{m}_i^{(p-1,p)}]$$

$$+\mathbb{E}\big[\big(-G_{ii}\mathrm{tr}\,\widetilde{B}G+(G_{ii}+T_i)\Upsilon\big)\mathfrak{m}_i^{(p-1,p)}\big]\,, \tag{5.26}$$

respectively,

$$\begin{aligned}\mathbb{E}[\mathfrak{n}_i^{(p,p)}]&=\mathbb{E}[K_i\mathfrak{n}_i^{(p-1,p)}]\\&=\mathbb{E}[T_i\mathfrak{n}_i^{(p-1,p)}]+\mathbb{E}\big[\big((b_iT_i+(\widetilde{B}G)_{ii})\mathrm{tr}\,G-(G_{ii}+T_i)\mathrm{tr}\,\widetilde{B}G\big)\mathfrak{n}_i^{(p-1,p)}\big]\,.\end{aligned} \tag{5.27}$$

Using the fact $\boldsymbol{e}_i^*R_i=-\boldsymbol{h}_i^*$ (*cf.*, (4.7)), we can write

$$\begin{aligned}(\widetilde{B}G)_{ii}&=\boldsymbol{e}_i^*R_i\widetilde{B}^{\langle i\rangle}R_iG\boldsymbol{e}_i=-\boldsymbol{h}_i^*\widetilde{B}^{\langle i\rangle}R_iG\boldsymbol{e}_i\\&=-\boldsymbol{h}_i^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i+\ell_i^2\boldsymbol{h}_i^*\widetilde{B}^{\langle i\rangle}(\boldsymbol{e}_i+\boldsymbol{h}_i)(\boldsymbol{e}_i+\boldsymbol{h}_i)^*G\boldsymbol{e}_i\\&=-S_i+\ell_i^2(b_ih_{ii}+\boldsymbol{h}_i^*\widetilde{B}^{\langle i\rangle}\boldsymbol{h}_i)(G_{ii}+T_i)=-\mathring{S}_i+\varepsilon_{i1}\,,\end{aligned} \tag{5.28}$$

where S_i and $\mathring{S}_i$ are defined in (4.10) and (5.5), respectively, ℓ_i is defined in (4.5) and

$$\varepsilon_{i1}:=\big((\ell_i^2-1)b_ih_{ii}+\ell_i^2\boldsymbol{h}_i^*\widetilde{B}^{\langle i\rangle}\boldsymbol{h}_i\big)G_{ii}+\ell_i^2\big(b_ih_{ii}+\boldsymbol{h}_i^*\widetilde{B}^{\langle i\rangle}\boldsymbol{h}_i\big)T_i\,. \tag{5.29}$$

With the aid of Lemma A.1, it is elementary to check

$$|h_{ii}|\prec\frac{1}{\sqrt{N}}\,,\qquad|\ell_i^2-1|\prec\frac{1}{\sqrt{N}}\,,\qquad|\boldsymbol{h}_i^*\widetilde{B}^{\langle i\rangle}\boldsymbol{h}_i|\prec\frac{1}{\sqrt{N}}\,, \tag{5.30}$$

where in the last inequality we also used the fact that $\mathrm{tr}\,\widetilde{B}^{\langle i\rangle}=\mathrm{tr}\,B=0$, under the convention (5.1). Applying the bounds in (5.17) and (5.30), it is easy to see that

$$|\varepsilon_{i1}|\prec\frac{1}{\sqrt{N}}\,. \tag{5.31}$$

Substituting (5.28) and (5.31) into the first term on the right hand side of (5.26), we have

$$\mathbb{E}[(\widetilde{B}G)_{ii}\mathrm{tr}\,G\mathfrak{m}_i^{(p-1,p)}]=-\mathbb{E}[\mathring{S}_i\mathrm{tr}\,G\mathfrak{m}_i^{(p-1,p)}]+\mathbb{E}[O_\prec(N^{-\frac12})\mathfrak{m}_i^{(p-1,p)}]\,, \tag{5.32}$$

where for the second term on the right hand side above we also used $\mathrm{tr}\,G=O_\prec(1)$; *cf.*, (5.20). We recall the definition of $\mathring{S}_i$ from (5.5) and rewrite

$$\mathring{S}_i=\sum_k^{(i)}\bar{g}_{ik}\frac{1}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i.$$

Hereafter, we use the notation $\sum_k^{(i)}$ to represent the sum over $k\in[\![1,N]\!]\setminus\{i\}$. Thus, the first term on the right of (5.32) is of the form $\mathbb{E}[\sum_k^{(i)}\bar{g}_{ik}\langle\cdots\rangle]$, where $\langle\cdots\rangle$ can be regarded as a function of the $\bar{g}_{ik}$'s and the g_{ik}'s. Recall the following integration by parts formula for complex centered Gaussian variables,

$$\int_{\mathbb{C}} \bar{g} f(g,\bar{g}) \mathrm{e}^{-\frac{|g|^2}{\sigma^2}} \mathrm{d}^2 g = \sigma^2 \int_{\mathbb{C}} \partial_g f(g,\bar{g}) \mathrm{e}^{-\frac{|g|^2}{\sigma^2}} \mathrm{d}^2 g , \tag{5.33}$$

for any differentiable function $f : \mathbb{C}^2 \to \mathbb{C}$. Applying (5.33) to the first term on the right of (5.32), we get

$$\begin{aligned} \mathbb{E}[\mathring{S}_i \operatorname{tr} G \mathfrak{m}_i^{(p-1,p)}] &= \frac{1}{N} \sum_k^{(i)} \mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i\|} \frac{\partial (\boldsymbol{e}_k^* \widetilde{B}^{\langle i \rangle} G \boldsymbol{e}_i)}{\partial g_{ik}} \operatorname{tr} G \mathfrak{m}_i^{(p-1,p)} \Big] \\ &+ \frac{1}{N} \sum_k^{(i)} \mathbb{E}\Big[\frac{\partial \|\boldsymbol{g}_i\|^{-1}}{\partial g_{ik}} \boldsymbol{e}_k^* \widetilde{B}^{\langle i \rangle} G \boldsymbol{e}_i \operatorname{tr} G \mathfrak{m}_i^{(p-1,p)} \Big] \\ &+ \frac{1}{N} \sum_k^{(i)} \mathbb{E}\Big[\frac{\boldsymbol{e}_k^* \widetilde{B}^{\langle i \rangle} G \boldsymbol{e}_i}{\|\boldsymbol{g}_i\|} \frac{\partial \operatorname{tr} G}{\partial g_{ik}} \mathfrak{m}_i^{(p-1,p)} \Big] \\ &+ \frac{p-1}{N} \sum_k^{(i)} \mathbb{E}\Big[\frac{\boldsymbol{e}_k^* \widetilde{B}^{\langle i \rangle} G \boldsymbol{e}_i}{\|\boldsymbol{g}_i\|} \operatorname{tr} G \frac{\partial P_i}{\partial g_{ik}} \mathfrak{m}_i^{(p-2,p)} \Big] \\ &+ \frac{p}{N} \sum_k^{(i)} \mathbb{E}\Big[\frac{\boldsymbol{e}_k^* \widetilde{B}^{\langle i \rangle} G \boldsymbol{e}_i}{\|\boldsymbol{g}_i\|} \operatorname{tr} G \frac{\partial \overline{P_i}}{\partial g_{ik}} \mathfrak{m}_i^{(p-1,p-1)} \Big]. \end{aligned} \tag{5.34}$$

Analogously, by $T_i = \mathring{T}_i + h_{ii} G_{ii}$, (5.5), the first bound in (5.17), the first bound in (5.30), and also (5.12), we can write the first term on the right hand side of (5.27) as

$$\mathbb{E}[T_i \mathfrak{n}_i^{(p-1,p)}] = \mathbb{E}[\mathring{T}_i \mathfrak{n}_i^{(p-1,p)}] + \mathbb{E}[O_{\prec}(N^{-\frac{1}{2}}) \mathfrak{n}_i^{(p-1,p)}] . \tag{5.35}$$

Similarly to (5.34), applying the integration by parts formula, we obtain

$$\begin{aligned} \mathbb{E}[\mathring{T}_i \mathfrak{n}_i^{(p-1,p)}] &= \frac{1}{N} \sum_k^{(i)} \mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i\|} \frac{\partial (\boldsymbol{e}_k^* G \boldsymbol{e}_i)}{\partial g_{ik}} \mathfrak{n}_i^{(p-1,p)} \Big] + \frac{1}{N} \sum_k^{(i)} \mathbb{E}\Big[\frac{\partial \|\boldsymbol{g}_i\|^{-1}}{\partial g_{ik}} \boldsymbol{e}_k^* G \boldsymbol{e}_i \mathfrak{n}_i^{(p-1,p)} \Big] \\ &+ \frac{p-1}{N} \sum_k^{(i)} \mathbb{E}\Big[\frac{\boldsymbol{e}_k^* G \boldsymbol{e}_i}{\|\boldsymbol{g}_i\|} \frac{\partial K_i}{\partial g_{ik}} \mathfrak{n}_i^{(p-2,p)} \Big] + \frac{p}{N} \sum_k^{(i)} \mathbb{E}\Big[\frac{\boldsymbol{e}_k^* G \boldsymbol{e}_i}{\|\boldsymbol{g}_i\|} \frac{\partial \overline{K_i}}{\partial g_{ik}} \mathfrak{n}_i^{(p-1,p-1)} \Big] . \end{aligned} \tag{5.36}$$

First, we consider the first term on the right side of (5.34). Recall ℓ_i from (4.5). For brevity, we set

$$c_i := \frac{\ell_i^2}{\|\boldsymbol{g}_i\|}. \tag{5.37}$$

It is elementary to derive that

$$\frac{\partial G}{\partial g_{ik}} = c_i \big(G \boldsymbol{e}_k (\boldsymbol{e}_i + \boldsymbol{h}_i)^* \widetilde{B}^{\langle i \rangle} R_i G + G R_i \widetilde{B}^{\langle i \rangle} \boldsymbol{e}_k (\boldsymbol{e}_i + \boldsymbol{h}_i)^* G \big) + \Delta_G(i,k) . \tag{5.38}$$

Here $\Delta_G(i,k)$ is a small remainder, defined as

$$\Delta_G(i,k) := -G\Delta_R(i,k)\widetilde{B}^{\langle i\rangle}R_iG - GR_i\widetilde{B}^{\langle i\rangle}\Delta_R(i,k)G, \tag{5.39}$$

where

$$\Delta_R(i,k) := \frac{\ell_i^2}{2\|\boldsymbol{g}_i\|^2}\bar{g}_{ik}\big(\boldsymbol{e}_i\boldsymbol{h}_i^* + \boldsymbol{h}_i\boldsymbol{e}_i^* + 2\boldsymbol{h}_i\boldsymbol{h}_i^*\big) - \frac{\ell_i^4}{2\|\boldsymbol{g}_i\|^3}g_{ii}\bar{g}_{ik}\big(\boldsymbol{e}_i + \boldsymbol{h}_i\big)\big(\boldsymbol{e}_i + \boldsymbol{h}_i\big)^*. \tag{5.40}$$

The $\Delta_G(i,k)$'s are irrelevant error terms. We handle quantities with $\Delta_G(i,k)$ separately in Appendix B.

Analogously to (7.55) of [6], using (5.38), we can get

$$\begin{aligned}\frac{1}{N}\sum_k^{(i)}\frac{\partial(\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i)}{\partial g_{ik}} &= -c_i\frac{1}{N}\sum_k^{(i)}\boldsymbol{e}_k^*\widetilde{B}^{(i)}G\boldsymbol{e}_k(b_iT_i + (\widetilde{B}G)_{ii})\\ &\quad + c_i\frac{1}{N}\sum_k^{(i)}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}GR_i\widetilde{B}^{\langle i\rangle}\boldsymbol{e}_k(G_{ii} + T_i)\\ &\quad + \frac{1}{N}\sum_k^{(i)}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}\Delta_G(i,k)\boldsymbol{e}_i\,. \end{aligned}\tag{5.41}$$

Note that T_i naturally appears in the first term of (5.34) after integrating by parts the $\mathring{S}_i$ term. This explains why we need to study the high moments of K_i to get another equation. Now, we claim that

$$\frac{1}{N}\sum_k^{(i)}\boldsymbol{e}_k^*\widetilde{B}^{(i)}G\boldsymbol{e}_k = \operatorname{tr}\widetilde{B}G + O_\prec(\Pi_i^2)\,, \qquad \frac{1}{N}\sum_k^{(i)}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}GR_i\widetilde{B}^{\langle i\rangle}\boldsymbol{e}_k = \operatorname{tr}\widetilde{B}G\widetilde{B} + O_\prec(\Pi_i^2)\,, \tag{5.42}$$

with Π_i given in (5.11). We state the proof for the first estimate in (5.42). Note that

$$\frac{1}{N}\sum_k^{(i)}\boldsymbol{e}_k^*\widetilde{B}^{(i)}G\boldsymbol{e}_k = \operatorname{tr}\widetilde{B}^{\langle i\rangle}G - \frac{1}{N}(\widetilde{B}^{\langle i\rangle}G)_{ii} = \operatorname{tr}\widetilde{B}^{\langle i\rangle}G + O_\prec(\frac{1}{N})\,, \tag{5.43}$$

where the last step follows from the identity $(\widetilde{B}^{\langle i\rangle}G)_{ii} = b_iG_{ii}$ and (5.17). Then, using that $\widetilde{B}^{\langle i\rangle} = R_i\widetilde{B}R_i$ and $R_i = I - \boldsymbol{r}_i\boldsymbol{r}_i^*$ (*cf.*, (4.2)), we see that

$$\operatorname{tr}\widetilde{B}G - \operatorname{tr}\widetilde{B}^{\langle i\rangle}G = \operatorname{tr}\widetilde{B}G - \operatorname{tr}R_i\widetilde{B}R_iG = \frac{1}{N}\boldsymbol{r}_i^*\widetilde{B}G\boldsymbol{r}_i + \frac{1}{N}\boldsymbol{r}_i^*G\widetilde{B}\boldsymbol{r}_i - \frac{1}{N}\boldsymbol{r}_i^*\widetilde{B}\boldsymbol{r}_i\boldsymbol{r}_i^*G\boldsymbol{r}_i\,.$$

Using (4.6), $\ell_i = 1 + O_\prec(\frac{1}{\sqrt{N}})$ and $\|\boldsymbol{r}_i^*\widetilde{B}\| \lesssim 1$, we get by Cauchy-Schwarz that

$$\big|\boldsymbol{r}_i^*\widetilde{B}G\boldsymbol{r}_i\big| \lesssim \Big(\|G\boldsymbol{e}_i\|^2+\|G\boldsymbol{h}_i\|^2\Big)^{\frac12} = \Big(\frac{\mathrm{Im}\,(G_{ii}+\boldsymbol{h}_i^*G\boldsymbol{h}_i)}{\eta}\Big)^{\frac12} = \Big(\frac{\mathrm{Im}\,(G_{ii}+\mathcal{G}_{ii})}{\eta}\Big)^{\frac12},$$

with $\mathcal{G}$ given in (5.3), where in the last step we used

$$\boldsymbol{h}_i^*G\boldsymbol{h}_i = \boldsymbol{u}_i^*G\boldsymbol{u}_i = \boldsymbol{e}_i^*U^*GU\boldsymbol{e}_i = \mathcal{G}_{ii} \tag{5.44}$$

and the identities $|G|^2 = \frac{1}{\eta}\mathrm{Im}\,G$ and $|\mathcal{G}|^2 = \frac{1}{\eta}\mathrm{Im}\,\mathcal{G}$. Similarly, we have

$$\big|\boldsymbol{r}_i^*G\widetilde{B}\boldsymbol{r}_i\big| \lesssim \Big(\frac{\mathrm{Im}\,(G_{ii}+\mathcal{G}_{ii})}{\eta}\Big)^{\frac12}, \qquad \big|\boldsymbol{r}_i^*G\boldsymbol{r}_i\big| \lesssim \Big(\frac{\mathrm{Im}\,(G_{ii}+\mathcal{G}_{ii})}{\eta}\Big)^{\frac12}.$$

Hence, we have

$$\big|\mathrm{tr}\,\widetilde{B}G - \mathrm{tr}\,\widetilde{B}^{\langle i\rangle}G\big| \lesssim \frac{1}{N}\Big(\frac{\mathrm{Im}\,(G_{ii}+\mathcal{G}_{ii})}{\eta}\Big)^{\frac12} \lesssim \frac{\mathrm{Im}\,(G_{ii}+\mathcal{G}_{ii})}{N\eta} = O_\prec(\Pi_i^2), \tag{5.45}$$

where in the second step, we used the fact $\mathrm{Im}\,G_{ii}, \mathrm{Im}\,\mathcal{G}_{ii} \gtrsim \eta$. Combining (5.43) with (5.45) we obtain the first estimate of (5.42). The second estimate in (5.42) is proved in the same way.

Hence, using (5.42) and the first estimate in (B.1), we obtain from (5.41) that

$$\frac{1}{N}\sum_k^{(i)}\frac{\partial(\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i)}{\partial g_{ik}} = -c_i\mathrm{tr}\,\widetilde{B}G\big(b_iT_i+(\widetilde{B}G)_{ii}\big)+c_i\mathrm{tr}\,\widetilde{B}G\widetilde{B}\big(G_{ii}+T_i\big)+O_\prec(\Pi_i^2)\,. \tag{5.46}$$

Analogously, we can show that

$$\frac{1}{N}\sum_k^{(i)}\frac{\partial(\boldsymbol{e}_k^*G\boldsymbol{e}_i)}{\partial g_{ik}} = -c_i\mathrm{tr}\,G\big(b_iT_i+(\widetilde{B}G)_{ii}\big)+c_i\mathrm{tr}\,\widetilde{B}G\big(G_{ii}+T_i\big)+O_\prec(\Pi_i^2)\,. \tag{5.47}$$

Using (5.27), (5.35), (5.36) and (5.47) and the estimate $\frac{c_i}{\|\boldsymbol{g}_i\|} = 1+O_\prec(\frac{1}{\sqrt{N}})$, we obtain

$$\begin{aligned}\mathbb{E}[\mathfrak{n}_i^{(p,p)}] &= \mathbb{E}\Big[O_\prec(\Psi)\mathfrak{n}_i^{(p-1,p)}\Big]+\frac{1}{N}\sum_k^{(i)}\mathbb{E}\Big[\frac{\partial\|\boldsymbol{g}_i\|^{-1}}{\partial g_{ik}}\boldsymbol{e}_k^*G\boldsymbol{e}_i\mathfrak{n}_i^{(p-1,p)}\Big]\\ &\quad+\frac{p-1}{N}\sum_k^{(i)}\mathbb{E}\Big[\frac{\boldsymbol{e}_k^*G\boldsymbol{e}_i}{\|\boldsymbol{g}_i\|}\frac{\partial K_i}{\partial g_{ik}}\mathfrak{n}_i^{(p-2,p)}\Big]+\frac{p}{N}\sum_k^{(i)}\mathbb{E}\Big[\frac{\boldsymbol{e}_k^*G\boldsymbol{e}_i}{\|\boldsymbol{g}_i\|}\frac{\partial\overline{K_i}}{\partial g_{ik}}\mathfrak{n}_i^{(p-1,p-1)}\Big].\end{aligned} \tag{5.48}$$

Then, combining (5.46) with (5.47), we obtain

$$\frac{1}{N}\sum_k^{(i)}\frac{\partial(\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i)}{\partial g_{ik}}\mathrm{tr}\,G = -c_i(G_{ii}+T_i)\big(\mathrm{tr}\,\widetilde{B}G-\Upsilon\big)+\frac{1}{N}\sum_k^{(i)}\frac{\partial(\boldsymbol{e}_k^*G\boldsymbol{e}_i)}{\partial g_{ik}}\mathrm{tr}\,\widetilde{B}G+O_\prec(\Pi_i^2)$$

$$= -c_i(G_{ii}+T_i)\big(\mathrm{tr}\,\widetilde{B}G-\Upsilon\big)+\mathring{T}_i\mathrm{tr}\,\widetilde{B}G+\Big(\frac{1}{N}\sum_k^{(i)}\frac{\partial(\boldsymbol{e}_k^*G\boldsymbol{e}_i)}{\partial g_{ik}}-\mathring{T}_i\Big)\mathrm{tr}\,\widetilde{B}G+O_\prec(\Pi_i^2)\,. \tag{5.49}$$

Recall the definition of c_i from (5.37). It is elementary to check that

$$c_i=\|\boldsymbol{g}_i\|-h_{ii}-\big(\|\boldsymbol{g}_i\|^2-1\big)+O_\prec(\frac{1}{N})\,. \tag{5.50}$$

Plugging (5.50) into (5.49) and also using the second equation in (5.5), we can write

$$\begin{aligned}\frac{1}{N}\sum_k^{(i)}\frac{\partial(\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i)}{\partial g_{ik}}\mathrm{tr}\,G&=-\|\boldsymbol{g}_i\|\big(G_{ii}\mathrm{tr}\,\widetilde{B}G-(G_{ii}+T_i)\Upsilon\big)\\&\quad+\Big(\frac{1}{N}\sum_k^{(i)}\frac{\partial(\boldsymbol{e}_k^*G\boldsymbol{e}_i)}{\partial g_{ik}}-\|\boldsymbol{g}_i\|\mathring{T}_i\Big)\mathrm{tr}\,\widetilde{B}G+\varepsilon_{i2}+O_\prec(\Pi_i^2),\end{aligned} \tag{5.51}$$

where ε_{i2} collects irrelevant terms

$$\begin{aligned}\varepsilon_{i2}:=&\big(\|\boldsymbol{g}_i\|-c_i\big)\big(G_{ii}\mathrm{tr}\,\widetilde{B}G-(G_{ii}+T_i)\Upsilon\big)+\big(\|\boldsymbol{g}_i\|\mathring{T}_i-c_iT_i\big)\mathrm{tr}\,\widetilde{B}G\\=&\big(\|\boldsymbol{g}_i\|^2-1\big)G_{ii}\mathrm{tr}\,\widetilde{B}G-\big(h_{ii}+\big(\|\boldsymbol{g}_i\|^2-1\big)\big)(G_{ii}+T_i)\Upsilon\\&+\big(h_{ii}+\big(\|\boldsymbol{g}_i\|^2-1\big)\big)T_i\mathrm{tr}\,\widetilde{B}G+O_\prec(\frac{1}{N})\,.\end{aligned} \tag{5.52}$$

From the estimates $|h_{ii}|\prec\frac{1}{\sqrt{N}}$, $\|\boldsymbol{g}_i\|=1+O_\prec(\frac{1}{\sqrt{N}})$, (5.17) and the observation that the tracial quantities are $O_\prec(1)$, we see that

$$\varepsilon_{i2}=O_\prec(\frac{1}{\sqrt{N}})\,. \tag{5.53}$$

Combining (5.26), (5.28), (5.34) and (5.51), we have

$$\begin{aligned}\mathbb{E}[\mathfrak{m}_i^{(p,p)}]&=-\mathbb{E}[(\mathring{S}_i+\varepsilon_{i1})\mathrm{tr}\,G\mathfrak{m}_i^{(p-1,p)}]+\mathbb{E}\big[\big(-G_{ii}\mathrm{tr}\,\widetilde{B}G+(G_{ii}+T_i)\Upsilon\big)\mathfrak{m}_i^{(p-1,p)}\big]\\&=\mathbb{E}\Big[\Big(\mathring{T}_i-\frac{1}{\|\boldsymbol{g}_i\|}\frac{1}{N}\sum_k^{(i)}\frac{\partial(\boldsymbol{e}_k^*G\boldsymbol{e}_i)}{\partial g_{ik}}\Big)\mathrm{tr}\,\widetilde{B}G\mathfrak{m}_i^{(p-1,p)}\Big]\\&\quad-\frac{1}{N}\sum_k^{(i)}\mathbb{E}\Big[\frac{\partial\|\boldsymbol{g}_i\|^{-1}}{\partial g_{ik}}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i\mathrm{tr}\,G\mathfrak{m}_i^{(p-1,p)}\Big]\\&\quad-\frac{1}{N}\sum_k^{(i)}\mathbb{E}\Big[\frac{\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i}{\|\boldsymbol{g}_i\|}\frac{\partial\mathrm{tr}\,G}{\partial g_{ik}}\mathfrak{m}_i^{(p-1,p)}\Big]\\&\quad-\frac{p-1}{N}\sum_k^{(i)}\mathbb{E}\Big[\frac{\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i}{\|\boldsymbol{g}_i\|}\mathrm{tr}\,G\frac{\partial P_i}{\partial g_{ik}}\mathfrak{m}_i^{(p-2,p)}\Big]\end{aligned}$$

$$
\begin{aligned}
&- \frac{p}{N} \sum_{k}^{(i)} \mathbb{E}\Big[\frac{\boldsymbol{e}_k^* \widetilde{B}^{\langle i \rangle} G \boldsymbol{e}_i}{\|\boldsymbol{g}_i\|} \operatorname{tr} G \frac{\partial \overline{P_i}}{\partial g_{ik}} \mathfrak{m}_i^{(p-1,p-1)} \Big] \\
&+ \mathbb{E}\Big[\Big(\varepsilon_{i1} \operatorname{tr} G - \frac{1}{\|\boldsymbol{g}_i\|} \varepsilon_{i2} + O_{\prec}(\Pi_i^2) \Big) \mathfrak{m}_i^{(p-1,p)} \Big]. \qquad (5.54)
\end{aligned}
$$

For the first term on the right of (5.54), analogously to (5.36), applying (5.33) to the $\mathring{T}_i$-term, we get

$$
\begin{aligned}
&\mathbb{E}\Big[\Big(\mathring{T}_i - \frac{1}{\|\boldsymbol{g}_i\|} \frac{1}{N} \sum_{k}^{(i)} \frac{\partial (\boldsymbol{e}_k^* G \boldsymbol{e}_i)}{\partial g_{ik}} \Big) \operatorname{tr} \widetilde{B} G \mathfrak{m}_i^{(p-1,p)} \Big] \\
&= \frac{1}{N} \sum_{k}^{(i)} \mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i\|} \frac{\partial \operatorname{tr} \widetilde{B} G}{\partial g_{ik}} \boldsymbol{e}_k^* G \boldsymbol{e}_i \operatorname{tr} \widetilde{B} G \mathfrak{m}_i^{(p-1,p)} \Big] + \frac{1}{N} \sum_{k}^{(i)} \mathbb{E}\Big[\frac{\partial \|\boldsymbol{g}_i\|^{-1}}{\partial g_{ik}} \boldsymbol{e}_k^* G \boldsymbol{e}_i \operatorname{tr} \widetilde{B} G \mathfrak{m}_i^{(p-1,p)} \Big] \\
&\quad + \frac{p-1}{N} \sum_{k}^{(i)} \mathbb{E}\Big[\frac{\boldsymbol{e}_k^* G \boldsymbol{e}_i}{\|\boldsymbol{g}_i\|} \frac{\partial P_i}{\partial g_{ik}} \operatorname{tr} \widetilde{B} G \mathfrak{m}_i^{(p-2,p)} \Big] + \frac{p}{N} \sum_{k}^{(i)} \mathbb{E}\Big[\frac{\boldsymbol{e}_k^* G \boldsymbol{e}_i}{\|\boldsymbol{g}_i\|} \frac{\partial \overline{P_i}}{\partial g_{ik}} \operatorname{tr} \widetilde{B} G \mathfrak{m}_i^{(p-1,p-1)} \Big]. \qquad (5.55)
\end{aligned}
$$

Recall the estimates of ε_{i1} and ε_{i2} in (5.31) and (5.53), respectively, which implies that $|\varepsilon_{i1}| \prec \Psi$ and $|\varepsilon_{i2}| \prec \Psi$. Therefore, to show (5.24), it suffices to estimate the four last terms in the right side of (5.54), and all the terms on the right side of (5.55). Then, in order to show (5.25), it suffices to estimate the last three terms on the right side of (5.48). All these terms can be estimated based on the following lemma.

Lemma 5.3. *Suppose the assumptions in Proposition 5.1 hold. Set $X_i = I$ or $\widetilde{B}^{\langle i \rangle}$. Let Q be any (possibly random) diagonal matrix satisfying $\|Q\| \prec 1$ and $X = I$ or A. We have the following estimates*

$$
\begin{aligned}
&\frac{1}{N} \sum_{k}^{(i)} \frac{\partial \|\boldsymbol{g}_i\|^{-1}}{\partial g_{ik}} \boldsymbol{e}_k^* X_i G \boldsymbol{e}_i = O_{\prec}\big(\frac{1}{N}\big), \qquad \frac{1}{N} \sum_{k}^{(i)} \boldsymbol{e}_i^* X \frac{\partial G}{\partial g_{ik}} \boldsymbol{e}_i \boldsymbol{e}_k^* X_i G \boldsymbol{e}_i = O_{\prec}(\Pi_i^2), \\
&\frac{1}{N} \sum_{k}^{(i)} \frac{\partial T_i}{\partial g_{ik}} \boldsymbol{e}_k^* X_i G \boldsymbol{e}_i = O_{\prec}(\Pi_i^2), \qquad \frac{1}{N} \sum_{k}^{(i)} \operatorname{tr}\Big(Q X \frac{\partial G}{\partial g_{ik}} \Big) \boldsymbol{e}_k^* X_i G \boldsymbol{e}_i = O_{\prec}\big(\Psi^2 \Pi_i^2\big), \\
&\frac{1}{N} \sum_{k}^{(i)} \operatorname{tr}\Big(Q X \frac{\partial G}{\partial g_{ik}} \Big) \boldsymbol{e}_k^* X_i \mathring{\boldsymbol{g}}_i = O_{\prec}\big(\Psi^2 \Pi_i^2\big). \qquad (5.56)
\end{aligned}
$$

In addition, the same estimates hold if we replace $\frac{\partial G}{\partial g_{ik}}$ and $\frac{\partial T_i}{\partial g_{ik}}$ by their complex conjugates $\frac{\partial \overline{G}}{\partial g_{ik}}$ and $\frac{\partial \overline{T}_i}{\partial g_{ik}}$ in the last four equations above.

The proof of Lemma 5.3 is postponed to Appendix B. We also remark here that last equation in (5.56) will not be used in the remaining proof of Lemma 5.2, but it will be used in Section 6. With the aid of Lemma 5.3, the remaining proof of Lemma 5.2 is the same as the counterpart to the proof of Lemma 7.3 in [6]. The only difference is that we use the improved bounds in Lemma 5.3 instead of those in Lemma 7.4 in [6]. Specifically, the estimates for the second term of (5.48), the second term of (5.54), and the second term of (5.55) follow from the first equation in (5.56). The third term of (5.54) and the first term of (5.55) can be estimated by the fourth equation in (5.56), after writing $\operatorname{tr}\widetilde{B}G = 1 - \operatorname{tr}(A-z)G$. All the other terms have $\frac{\partial K_i}{\partial g_{ik}}$ and $\frac{\partial P_i}{\partial g_{ik}}$ or their complex conjugate involved. Recall the definitions in (4.14) and (4.15), and also the first equation in (5.9). Then, by the chain rule, we see that all terms in (5.48), (5.54) and (5.55), with $\frac{\partial K_i}{\partial g_{ik}}$ and $\frac{\partial P_i}{\partial g_{ik}}$ or their complex conjugate counterparts involved, can be estimated by combining the second to the fourth equations in (5.56). This completes the proof of Lemma 5.2. $\square$

With Lemma 5.2, we can complete the proof of Proposition 5.1. The proof is nearly the same as that for Theorem 7.2 in [6]. For the convenience of the reader, we sketch it below.

Fix $z \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$. Using Young's inequality, we obtain from (5.24) that for any given (small) $\varepsilon > 0$,

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

Since $\varepsilon > 0$ was arbitrary, this implies the first bound in (5.22). The second one then follows from (5.25) in the same manner. By Markov's inequality, we get (5.14).

Next, we show how (5.15) and (5.16) follow from (5.14) and the assumption (5.13). To this end, we first prove the following crude bound

$$\Lambda_T(z) \prec N^{-\frac{\gamma}{4}}. \tag{5.57}$$

From the definition in (4.15), we can rewrite the second estimate in (5.14) as

$$(1 + b_i \operatorname{tr} G - \operatorname{tr}(\widetilde{B}G))T_i = G_{ii}\operatorname{tr}(\widetilde{B}G) - (\widetilde{B}G)_{ii}\operatorname{tr} G + O_\prec(\Psi). \tag{5.58}$$

Using the identity

$$(\widetilde{B}G)_{ii} = 1 - (a_i - z)G_{ii}(z), \tag{5.59}$$

and approximate G_{ii} by $(a_i - \omega_B)^{-1}$, we get from (5.13) and (3.2) that

$$(\widetilde{B}G)_{ii} = \frac{z - \omega_B}{a_i - \omega_B} + O_\prec(N^{-\frac{\gamma}{4}}). \tag{5.60}$$

We also recall the estimates of the tracial quantities in (5.20) under the assumption (5.13). Plugging (5.60), (5.20) and the first bound in the assumption (5.13) into (5.58), we get

$$\big(1+(b_i-z+\omega_B)m_{\mu_A\boxplus\mu_B}+O_\prec(N^{-\frac{\gamma}{4}})\big)T_i=O_\prec(N^{-\frac{\gamma}{4}})+O_\prec(\Psi)=O_\prec(N^{-\frac{\gamma}{4}})\,, \tag{5.61}$$

where in the last step we used that $\Psi \le N^{-\frac{\gamma}{2}}$ for all $\eta \ge \eta_{\mathrm{m}}$. From the second line in (2.11), we note that

$$1+(b_i-z+\omega_B)m_{\mu_A\boxplus\mu_B}=m_{\mu_A\boxplus\mu_B}\Big(\frac{1}{m_{\mu_A\boxplus\mu_B}}+b_i-z+\omega_B\Big)=m_{\mu_A\boxplus\mu_B}(b_i-\omega_A)\,.$$

Using (3.2) and $\|A\|,\|B\|\le C$, we get $|m_{\mu_A\boxplus\mu_B}(b_i-\omega_A)|\gtrsim 1$. This together with (5.61) implies (5.57).

To prove (5.15), we recall the definition of P_i in (4.14), which implies that

$$\frac{1}{N}\sum_{i=1}^N (G_{ii}+T_i)\Upsilon=\frac{1}{N}\sum_{i=1}^N P_i=O_\prec(\Psi)\,. \tag{5.62}$$

Using the facts $\frac{1}{N}\sum_{i=1}^N G_{ii}=m_{\mu_A\boxplus\mu_B}+O_\prec(N^{-\frac{\gamma}{4}})$ (*cf.*, (5.20)), and $\frac{1}{N}\sum_{i=1}^N T_i=O_\prec(N^{-\frac{\gamma}{4}})$, and also $|m_{\mu_A\boxplus\mu_B}|\gtrsim 1$, we get (5.15) from (5.62).

Then, combining (5.15) with the first estimate in (5.14), we get

$$(\widetilde{B}G)_{ii}\mathrm{tr}\,G-G_{ii}\mathrm{tr}\,\widetilde{B}G=O_\prec(\Psi)\,. \tag{5.63}$$

Applying the identity (5.59) and the definition of ω_B^c, we can rewrite (5.63) as

$$\big((a_i-\omega_B^c)G_{ii}-1\big)\mathrm{tr}\,G=O_\prec(\Psi)\,.$$

As shown above that $|\mathrm{tr}\,G|\gtrsim 1$ with high probability under the assumption (5.13), we get $(a_i-\omega_B^c)G_{ii}-1=O_\prec(\Psi)$. By (5.21) and (3.2), we also note that $|a_i-\omega_B^c|\gtrsim 1$ with high probability. This further implies the first estimate in (5.16).

Finally, plugging (5.63) back to (5.58), we can improve the right hand side of (5.61) to $O_\prec(\Psi)$. Then the second estimate in (5.16) follows. This completes the proof of Proposition 5.1. □

6. Rough fluctuation averaging for general linear combinations

In this section, we prove a rough fluctuation averaging estimate for the basic quantities Q_i defined in (4.11). From (5.63), we see that

$$|Q_i(z)|\prec\Psi\,,\qquad i\in[\![1,N]\!]\,,\qquad z\in\mathcal{D}_\tau(\eta_{\mathrm{m}},\eta_{\mathrm{M}})\,, \tag{6.1}$$

if the assumptions of Proposition 5.1 hold.

Recall the definition of the control parameters Π and Π_i in (4.9) and (5.11), respectively. The following proposition states that the average of the Q_i's is typically smaller than an individual Q_i.

Proposition 6.1. *Fix a $z \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$. Suppose that the assumptions of Proposition 5.1 hold. Set $X_i = I$ or $\widetilde{B}^{\langle i \rangle}$. Let $d_1, \ldots, d_N \in \mathbb{C}$ be possibly H-dependent quantities satisfying $\max_j |d_j| \prec 1$. Assume that they depend only weakly on the randomness in the sense that the following hold, for all $i, j \in [\![1, N]\!]$,*

$$\frac{1}{N} \sum_k^{(i)} \frac{\partial d_j}{\partial g_{ik}} \boldsymbol{e}_k^* X_i G \boldsymbol{e}_i = O_\prec\big(\Psi^2 \Pi_i^2\big)\,, \qquad \frac{1}{N} \sum_k^{(i)} \frac{\partial d_j}{\partial g_{ik}} \boldsymbol{e}_k^* X_i \mathring{\boldsymbol{g}}_i = O_\prec\big(\Psi^2 \Pi_i^2\big)\,, \quad (6.2)$$

and the same bounds hold when the d_j's are replaced by their complex conjugates $\overline{d_j}$. Suppose that $\Pi(z) \prec \hat{\Pi}(z)$ for some deterministic and positive function $\hat{\Pi}(z)$ that satisfies $\frac{1}{\sqrt{N\sqrt{\eta}}} + \Psi^2 \prec \hat{\Pi} \prec \Psi$. Then,

$$\Big| \frac{1}{N} \sum_{i=1}^N d_i Q_i \Big| \prec \Psi \hat{\Pi}\,. \qquad (6.3)$$

We remark that whenever the d_j's are deterministic, (6.2) trivially holds. However, we will also need (6.3) with certain random d_j's that satisfy (6.2).

For any d_i's satisfying the assumption in Proposition 6.1, we introduce the notation

$$\mathfrak{m}^{(k,l)} := \Big(\frac{1}{N} \sum_{i=1}^N d_i Q_i \Big)^k \Big(\frac{1}{N} \sum_{i=1}^N \overline{d_i}\, \overline{Q_i} \Big)^l\,, \qquad k, l \in \mathbb{N}\,. \qquad (6.4)$$

Similarly to Lemma 5.2, it suffices to prove the following recursive moment estimate.

Lemma 6.2. *Fix a $z \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$. Suppose that the assumptions of Proposition 6.1 hold. Then, for any fixed integer $p \geq 1$, we have*

$$\mathbb{E}\big[\mathfrak{m}^{(p,p)}\big] = \mathbb{E}\big[O_\prec(\hat{\Pi}^2)\mathfrak{m}^{(p-1,p)}\big] + \mathbb{E}\big[O_\prec(\Psi^2\hat{\Pi}^2)\mathfrak{m}^{(p-2,p)}\big] + \mathbb{E}\big[O_\prec(\Psi^2\hat{\Pi}^2)\mathfrak{m}^{(p-1,p-1)}\big]. \qquad (6.5)$$

Proof of Proposition 6.1. Like the proof of (5.14) from Lemma 5.2, with Lemma 6.2, we can get (6.3) by applying Young's and Markov's inequalities. This completes the proof of Proposition 6.1. □

Proof of Lemma 6.2. We first claim that it suffices to prove the following statement: If $|\Upsilon(z)| \prec \hat{\Upsilon}(z)$ for any deterministic and positive function $\hat{\Upsilon}(z) \leq \Psi(z)$, then

$$\begin{aligned} \mathbb{E}\big[\mathfrak{m}^{(p,p)}\big] =& \mathbb{E}\big[(O_\prec(\hat{\Pi}^2) + O_\prec(\Psi\hat{\Upsilon}))\mathfrak{m}^{(p-1,p)}\big] + \mathbb{E}\big[O_\prec(\Psi^2\hat{\Pi}^2)\mathfrak{m}^{(p-2,p)}\big] \\ &+ \mathbb{E}\big[O_\prec(\Psi^2\hat{\Pi}^2)\mathfrak{m}^{(p-1,p-1)}\big]\,. \end{aligned} \qquad (6.6)$$

Indeed, the same as the proof of (5.14) from Lemma 5.2, we can again apply Young's inequality and Markov's inequality to get, for any d_i's satisfying the assumptions in Proposition 6.1, that (6.6) implies

$$\Big|\frac{1}{N}\sum_{i=1}^{N} d_i Q_i\Big| \prec \hat{\Pi}^2 + \Psi\hat{\Upsilon} + \Psi\hat{\Pi} \prec \Psi\hat{\Upsilon} + \Psi\hat{\Pi}\,, \tag{6.7}$$

where in the last step we used the assumption $\hat{\Pi} \prec \Psi$.

Next, recall from (5.10) that

$$\Upsilon = -\frac{1}{N}\sum_{i=1}^{N} a_i Q_i\,.$$

Choosing $d_i = a_i$ for all i, we get from (6.7)

$$|\Upsilon| \prec \Psi\hat{\Upsilon} + \Psi\hat{\Pi} \prec N^{-\frac{\gamma}{4}}\hat{\Upsilon} + \Psi\hat{\Pi}\,. \tag{6.8}$$

Using the right hand side of (6.8) as a new deterministic bound of Υ instead of the initial $\hat{\Upsilon}$ in (6.6), and perform the above argument iteratively, we can finally get

$$|\Upsilon| \prec \Psi\hat{\Pi}\,. \tag{6.9}$$

Hence, at the end, we can choose $\hat{\Upsilon} = \Psi\hat{\Pi}$ in (6.6) and get

$$\begin{aligned}\mathbb{E}\big[\mathfrak{m}^{(p,p)}\big] =& \mathbb{E}\big[(O_\prec(\hat{\Pi}^2) + O_\prec(\Psi^2\hat{\Pi}))\mathfrak{m}^{(p-1,p)}\big] + \mathbb{E}\big[O_\prec(\Psi^2\hat{\Pi}^2)\mathfrak{m}^{(p-2,p)}\big] \\ &+ \mathbb{E}\big[O_\prec(\Psi^2\hat{\Pi}^2)\mathfrak{m}^{(p-1,p-1)}\big].\end{aligned} \tag{6.10}$$

Observe that by the assumption $\frac{1}{\sqrt{N\sqrt{\eta}}} + \Psi^2 \prec \hat{\Pi}$, the $O_\prec(\Psi^2\hat{\Pi})$ term can be absorbed in the $O_\prec(\hat{\Pi}^2)$ term in (6.10). Hence, we conclude (6.5) from (6.6). Therefore, in the sequel, we will focus on proving (6.6).

Denote by $D := \mathrm{diag}(d_i)_{i=1}^N$. We first write

$$\frac{1}{N}\sum_{i=1}^{N} d_i Q_i = \frac{1}{N}\sum_{i=1}^{N}(\widetilde{B}G)_{ii}\big(d_i \mathrm{tr}\, G - \mathrm{tr}\, DG\big) = \frac{1}{N}\sum_{i=1}^{N}(\widetilde{B}G)_{ii}\mathrm{tr}\, G\tau_{i1}, \tag{6.11}$$

where we introduced the notation

$$\tau_{i1} := d_i - \frac{\mathrm{tr}\, DG}{\mathrm{tr}\, G}. \tag{6.12}$$

Similarly to the proof of (5.14), we approximate $(\widetilde{B}G)_{ii}$ by $-\mathring{S}_i$ (*cf.*, (5.28)), and then perform an integration by parts using (5.33) with respect to $\mathring{\boldsymbol{g}}_i$ in $\mathring{S}_i$. More specifically, we write

$$\mathbb{E}\big[\mathfrak{m}^{(p,p)}\big] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\Big[(\widetilde{B}G)_{ii}\mathrm{tr}\, G\tau_{i1}\mathfrak{m}^{(p-1,p)}\Big]$$

$$= -\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\Big[\mathring{S}_i \mathrm{tr}\, G\tau_{i1}\mathfrak{m}^{(p-1,p)}\Big] + \mathbb{E}\Big[\varepsilon_1 \mathfrak{m}^{(p-1,p)}\Big], \tag{6.13}$$

where we used the notation

$$\varepsilon_1 := \frac{1}{N}\sum_{i=1}^{N}\varepsilon_{i1}\mathrm{tr}\, G\tau_{i1}. \tag{6.14}$$

Here ε_{i1} is defined in (5.29). To ease the presentation, we further introduce the notation

$$\tau_{i2} := -\tau_{i1}\mathrm{tr}\, \widetilde{B}G. \tag{6.15}$$

Using assumption (5.13), (5.20), and also (3.2), one checks that $|\tau_{i1}| \prec 1$, $|\tau_{i2}| \prec 1$, for all $i \in [\![1, N]\!]$.

Analogously to (5.34), applying (5.33) to the first term on the right hand side of (6.13), we obtain

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\Big[\mathring{S}_i \mathrm{tr}\, G\tau_{i1}\mathfrak{m}^{(p-1,p)}\Big] &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{k}^{(i)}\mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i\|}\frac{\partial(\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i)}{\partial g_{ik}}\mathrm{tr}\, G\tau_{i1}\mathfrak{m}^{(p-1,p)}\Big] \\
&+ \frac{1}{N^2}\sum_{i=1}^{N}\sum_{k}^{(i)}\mathbb{E}\Big[\frac{\partial\|\boldsymbol{g}_i\|^{-1}}{\partial g_{ik}}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i \mathrm{tr}\, G\tau_{i1}\mathfrak{m}^{(p-1,p)}\Big] \\
&+ \frac{1}{N^2}\sum_{i=1}^{N}\sum_{k}^{(i)}\mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i\frac{\partial(\mathrm{tr}\, G\tau_{i1})}{\partial g_{ik}}\mathfrak{m}^{(p-1,p)}\Big] \\
&+ \frac{p-1}{N^2}\sum_{i=1}^{N}\sum_{k}^{(i)}\mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i \mathrm{tr}\, G\tau_{i1}\Big(\frac{1}{N}\sum_{j=1}^{N}\frac{\partial(d_jQ_j)}{\partial g_{ik}}\Big)\mathfrak{m}^{(p-2,p)}\Big] \\
&+ \frac{p}{N^2}\sum_{i=1}^{N}\sum_{k}^{(i)}\mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i \mathrm{tr}\, G\tau_{i1}\Big(\frac{1}{N}\sum_{j=1}^{N}\frac{\partial(\overline{d_jQ_j})}{\partial g_{ik}}\Big)\mathfrak{m}_i^{(p-1,p-1)}\Big].
\end{aligned} \tag{6.16}$$

First, we estimate the first term on the right hand side of (6.16). Using (5.51) and the bound

$$\frac{1}{N}\sum_{i=1}^{N}\Pi_i^2 \le 2\Pi^2,$$

we have

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{k}^{(i)}\frac{1}{\|\boldsymbol{g}_i\|}\frac{\partial(\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i)}{\partial g_{ik}}\mathrm{tr}\, G\tau_{i1} = -\frac{1}{N}\sum_{i=1}^{N}\big(G_{ii}\mathrm{tr}\, \widetilde{B}G - (G_{ii}+T_i)\Upsilon\big)\tau_{i1}$$

$$+\frac{1}{N^2}\sum_{i=1}^{N}\sum_{k}^{(i)}\Big(\mathring{T}_i-\frac{1}{\|\boldsymbol{g}_i\|}\frac{\partial(\boldsymbol{e}_k^*G\boldsymbol{e}_i)}{\partial g_{ik}}\Big)\tau_{i2}+\varepsilon_2+O_\prec(\Pi^2)\,,$$

where we have introduced

$$\varepsilon_2:=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\|\boldsymbol{g}_i\|}\tau_{i1}\varepsilon_{i2}\,; \tag{6.17}$$

see (5.52) for the definition of ε_{i2}. According to the definition in (6.12), we observe that

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\big(G_{ii}\mathrm{tr}\,\widetilde{B}G-(G_{ii}+T_i)\Upsilon\big)\tau_{i1}&=\frac{1}{N^2}\sum_{i=1}^{N}G_{ii}\tau_{i1}\big(\mathrm{tr}\,\widetilde{B}G-\Upsilon\big)-\frac{1}{N}\sum_{i=1}^{N}T_i\tau_{i1}\Upsilon\\&=O_\prec(\Psi\hat{\Upsilon})\,.\end{aligned} \tag{6.18}$$

Here in the last step we used the facts

$$\sum_{i=1}^{N}G_{ii}\tau_{i1}=0\,,\qquad\qquad \frac{1}{N}\sum_{i=1}^{N}T_i\tau_{i1}\Upsilon=O_\prec(\Psi\hat{\Upsilon})\,, \tag{6.19}$$

where the second estimate is implied by the second estimate in (5.16), and the assumption that $|\Upsilon|\prec\hat{\Upsilon}$.

Therefore, for the first term on the right hand side of (6.16), we have

$$\begin{aligned}&\frac{1}{N^2}\sum_{i=1}^{N}\sum_{k}^{(i)}\mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i\|}\frac{\partial(\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i)}{\partial g_{ik}}\mathrm{tr}\,G\tau_{i1}\mathfrak{m}^{(p-1,p)}\Big]\\&=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{k}^{(i)}\mathbb{E}\Big[\Big(\mathring{T}_i-\frac{1}{\|\boldsymbol{g}_i\|}\frac{\partial(\boldsymbol{e}_k^*G\boldsymbol{e}_i)}{\partial g_{ik}}\Big)\tau_{i2}\mathfrak{m}^{(p-1,p)}\Big]\\&\quad+\mathbb{E}\big[(\varepsilon_2+O_\prec(\Pi^2)+O_\prec(\Psi\hat{\Upsilon}))\mathfrak{m}^{(p-1,p)}\big]\\&=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{k}^{(i)}\mathbb{E}\Big[\frac{\partial\|\boldsymbol{g}_i\|^{-1}}{\partial g_{ik}}\boldsymbol{e}_k^*G\boldsymbol{e}_i\tau_{i2}\mathfrak{m}^{(p-1,p)}\Big]+\frac{1}{N^2}\sum_{i=1}^{N}\sum_{k}^{(i)}\mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i\|}\frac{\partial\tau_{i2}}{\partial g_{ik}}\boldsymbol{e}_k^*G\boldsymbol{e}_i\mathfrak{m}^{(p-1,p)}\Big]\\&\quad+\frac{p-1}{N^2}\sum_{i=1}^{N}\sum_{k}^{(i)}\mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k^*G\boldsymbol{e}_i\tau_{i2}\Big(\frac{1}{N}\sum_{j=1}^{N}\frac{\partial(d_jQ_j)}{\partial g_{ik}}\Big)\mathfrak{m}^{(p-2,p)}\Big]\\&\quad+\frac{p}{N^2}\sum_{i=1}^{N}\sum_{k}^{(i)}\mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k^*G\boldsymbol{e}_i\tau_{i2}\Big(\frac{1}{N}\sum_{j=1}^{N}\frac{\partial(\overline{d_jQ_j})}{\partial g_{ik}}\Big)\mathfrak{m}^{(p-1,p-1)}\Big]\\&\quad+\mathbb{E}\big[\big(\varepsilon_2+O_\prec(\Pi^2)+O_\prec(\Psi\hat{\Upsilon})\big)\mathfrak{m}^{(p-1,p)}\big],\end{aligned} \tag{6.20}$$

where the second equation is obtained analogously to (5.55), by writing $\mathring{T}_i = \sum_k^{(i)} \bar{g}_{ik} \boldsymbol{e}_k^* G \boldsymbol{e}_i / \|\boldsymbol{g}_i\|$ and performing integration by parts with respect to the g_{ik}'s.

According to (6.13), (6.16), and (6.20), it suffices to estimate the last term on the right side of (6.13), the last four terms on the right side of (6.16), and all the terms on the right side of (6.20). All the desired estimates can be derived from the following lemma.

Lemma 6.3. *Fix a $z \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$. Suppose that the assumptions of Proposition 6.1 hold, especially (6.2) holds for $d_1, \ldots, d_N$ in the definition (6.4). Let $\tilde{d}_1, \ldots, \tilde{d}_N \in \mathbb{C}$ be any (possibly random) numbers with the bound $\max_i |\tilde{d}_i| \prec 1$. Let Q be any (possibly random) diagonal matrix that satisfies $\|Q\| \prec 1$. Set $X = I$ or A, and set $X_i = I$ or $\widetilde{B}^{\langle i \rangle}$. Then we have*

$$\frac{1}{N^2} \sum_{i=1}^N \sum_k^{(i)} \tilde{d}_i \frac{\partial \|\boldsymbol{g}_i\|^{-1}}{\partial g_{ik}} \boldsymbol{e}_k^* X_i G \boldsymbol{e}_i = O_\prec\Big(\frac{1}{N}\Big)\,, \tag{6.21}$$

$$\frac{1}{N^2} \sum_{i=1}^N \sum_k^{(i)} \tilde{d}_i \mathrm{tr}\left(QX \frac{\partial G}{\partial g_{ik}}\right) \boldsymbol{e}_k^* X_i G \boldsymbol{e}_i = O_\prec(\Psi^2 \Pi^2)\,, \tag{6.22}$$

and the same estimate holds if we replace $\frac{\partial G}{\partial g_{ik}}$ by the complex conjugate $\frac{\partial \overline{G}}{\partial g_{ik}}$ in (6.22). Further, we have

$$\begin{aligned} \mathbb{E}\big[\varepsilon_j \mathfrak{m}^{(p-1,p)}\big] &= \mathbb{E}\big[O_\prec(\hat{\Pi}^2)\mathfrak{m}^{(p-1,p)}\big] \\ &\quad + \mathbb{E}\big[O_\prec(\Psi^2\hat{\Pi}^2)\mathfrak{m}^{(p-2,p)}\big] + \mathbb{E}\big[O_\prec(\Psi^2\hat{\Pi}^2)\mathfrak{m}^{(p-1,p-1)}\big]\,, \qquad j = 1, 2. \end{aligned} \tag{6.23}$$

We postpone the proof of Lemma 6.3 for a moment and continue with the proof of Lemma 6.2 instead.

The second term of (6.16) and the first term of (6.20) are directly estimated by (6.21). Using the definition of τ_{i1} in (6.12) and of τ_{i2} in (6.15), the boundedness of the tracial quantities (*cf.*, (5.20)), and the chain rule, we get the estimate on the third term of (6.16) and the second term of (6.20), using (6.22) and the assumption (6.2). For the last two terms of (6.16), and the third and fourth terms of (6.20), we note that

$$\begin{aligned} \frac{1}{N} \sum_{j=1}^N d_j Q_j &= \mathrm{tr}\, D\widetilde{B}G \,\mathrm{tr}\, G - \mathrm{tr}\, \widetilde{B}G \,\mathrm{tr}\, DG \\ &= \mathrm{tr}\, D \,\mathrm{tr}\, G - \mathrm{tr}\, DG - \mathrm{tr}\, DAG \,\mathrm{tr}\, G + \mathrm{tr}\, AG \,\mathrm{tr}\, DG\,, \end{aligned}$$

where in the last step we used the first identity of (5.9). Hence, by the chain rule, the fourth term of (6.16) and the third term of (6.20) are estimated with the aid of (6.22) and (6.2). The last term of (6.16) and the fourth term of (6.20) can be estimated analogously.

Finally, the estimates of the second term of (6.13) and the last term of (6.20) are given by (6.23). Thus we conclude the proof of Lemma 6.2. □

Proof of Lemma 6.3. Note that (6.21) and (6.22) follow from the first and the second last estimates in (5.56), respectively, by averaging over the index i. Hence, it suffices to prove (6.23). Recall the definition of ε_1 from (6.14) and of ε_2 from (6.17).

We first consider $\mathbb{E}[\varepsilon_1 \mathfrak{m}^{(p-1,p)}]$. Recall the definition of ε_{i1} from (5.29). Using (5.15), (5.16), the first bound in (5.17), and (5.30), we have

$$\varepsilon_{i1} = \frac{\boldsymbol{h}_i^* \widetilde{B}^{\langle i\rangle} \boldsymbol{h}_i}{a_i - \omega_B^c} + O_\prec\Big(\frac{\Psi}{\sqrt{N}}\Big) = \frac{\mathring{\boldsymbol{h}}_i^* \widetilde{B}^{\langle i\rangle} \mathring{\boldsymbol{h}}_i}{a_i - \omega_B^c} + O_\prec(\hat{\Pi}^2)\,. \tag{6.24}$$

Here the last step follows from the assumption $\frac{1}{N\sqrt{\eta}} \prec \hat{\Pi}^2$, and that $\boldsymbol{h}_i = \mathring{\boldsymbol{h}}_i + \frac{g_{ii}}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_i$ with

$$|g_{ii}| \prec \frac{1}{\sqrt{N}}\,, \qquad\qquad \mathring{\boldsymbol{h}}_i^* \widetilde{B}^{\langle i\rangle} \boldsymbol{e}_i = b_i \mathring{\boldsymbol{h}}_i^* \boldsymbol{e}_i = 0\,.$$

Hence, by the definition of ε_1 in (6.14), we have

$$\varepsilon_1 = \frac{1}{N}\sum_{i=1}^N \mathring{\boldsymbol{h}}_i^* \widetilde{B}^{\langle i\rangle} \mathring{\boldsymbol{h}}_i \frac{d_i \mathrm{tr}\, G - \mathrm{tr}\, DG}{a_i - \omega_B^c} + O_\prec(\hat{\Pi}^2) = \frac{1}{N}\sum_{i=1}^N \mathring{\boldsymbol{h}}_i^* \widetilde{B}^{\langle i\rangle} \mathring{\boldsymbol{h}}_i \tau_{i3} + O_\prec(\hat{\Pi}^2)\,,$$

where we introduced the notation

$$\tau_{i3} := \frac{d_i \mathrm{tr}\, G - \mathrm{tr}\, DG}{a_i - \omega_B^c}\,.$$

Using the integration by parts formula (5.33), we obtain

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \mathbb{E}\big[\mathring{\boldsymbol{h}}_i^* \widetilde{B}^{\langle i\rangle} \mathring{\boldsymbol{h}}_i \tau_{i3} \mathfrak{m}^{(p-1,p)}\big] &= \frac{1}{N}\sum_{i=1}^N \sum_k^{(i)} \mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i\|^2} \bar{g}_{ik} \boldsymbol{e}_k^* \widetilde{B}^{\langle i\rangle} \mathring{\boldsymbol{g}}_i \tau_{i3} \mathfrak{m}^{(p-1,p)}\Big] \\ &= \frac{1}{N^2}\sum_{i=1}^N \sum_k^{(i)} \mathbb{E}\Big[\frac{\partial\big(\|\boldsymbol{g}_i\|^{-2} \boldsymbol{e}_k^* \widetilde{B}^{\langle i\rangle} \mathring{\boldsymbol{g}}_i \tau_{i3} \mathfrak{m}^{(p-1,p)}\big)}{\partial g_{ik}}\Big].\end{aligned} \tag{6.25}$$

Note that

$$\begin{aligned}\frac{\partial\big(\|\boldsymbol{g}_i\|^{-2} \boldsymbol{e}_k^* \widetilde{B}^{\langle i\rangle} \mathring{\boldsymbol{g}}_i \tau_{i3} \mathfrak{m}^{(p-1,p)}\big)}{\partial g_{ik}} &= \frac{\partial \|\boldsymbol{g}_i\|^{-2}}{\partial g_{ik}} \boldsymbol{e}_k^* \widetilde{B}^{\langle i\rangle} \mathring{\boldsymbol{g}}_i \tau_{i3} \mathfrak{m}^{(p-1,p)} \\ &+ \|\boldsymbol{g}_i\|^{-2} \boldsymbol{e}_k^* \widetilde{B}^{\langle i\rangle} \boldsymbol{e}_k \tau_{i3} \mathfrak{m}^{(p-1,p)} + \|\boldsymbol{g}_i\|^{-2} \boldsymbol{e}_k^* \widetilde{B}^{\langle i\rangle} \mathring{\boldsymbol{g}}_i \frac{\partial \tau_{i3}}{\partial g_{ik}} \mathfrak{m}^{(p-1,p)}\end{aligned}$$

$$+ (p-1)\|\boldsymbol{g}_i\|^{-2}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}\mathring{\boldsymbol{g}}_i\tau_{i3}\Big(\frac{1}{N}\sum_{j=1}^N\frac{\partial(d_jQ_j)}{\partial g_{ik}}\Big)\mathfrak{m}^{(p-2,p)}$$

$$+ p\|\boldsymbol{g}_i\|^{-2}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}\mathring{\boldsymbol{g}}_i\tau_{i3}\Big(\frac{1}{N}\sum_{j=1}^N\frac{\partial(\overline{d_jQ_j})}{\partial g_{ik}}\Big)\mathfrak{m}^{(p-1,p-1)}\,. \tag{6.26}$$

Notice that $\frac{\partial\|\boldsymbol{g}_i\|^{-2}}{\partial g_{ik}} = -\|\boldsymbol{g}_i\|^{-4}\bar{g}_{ik}$ and that $\tau_{i3} = O_\prec(1)$. In addition, we also have that

$$\sum_k^{(i)}\bar{g}_{ik}\boldsymbol{e}_k = \mathring{\boldsymbol{g}}_i^*\,, \qquad\qquad \sum_k^{(i)}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}\boldsymbol{e}_k = \mathrm{Tr}B - b_i = b_i\,.$$

Denoting by $\tilde{d}_1,\ldots,\tilde{d}_N\in\mathbb{C}$ generic (possibly random) numbers with $\max_i|\tilde{d}_i|\prec 1$, we see that the contributions from the first two terms on the right side of (6.26) to (6.25) follow from the estimates

$$\frac{1}{N^2}\sum_{i=1}^N\tilde{d}_i\mathring{\boldsymbol{g}}_i^*\widetilde{B}^{\langle i\rangle}\mathring{\boldsymbol{g}}_i = O_\prec(\frac{1}{N})\,, \qquad\qquad \frac{1}{N^2}\sum_{i=1}^N\tilde{d}_ib_i\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}\boldsymbol{e}_k = O_\prec(\frac{1}{N})\,.$$

Here $\tilde{d}_i$ includes τ_{i3} and an appropriate power of $\|\boldsymbol{g}_i\|$. In addition, for the estimate of the remaining terms in (6.26), we claim that, for $X_i = I, \widetilde{B}^{\langle i\rangle}$,

$$\frac{1}{N^2}\sum_{i=1}^N\sum_k^{(i)}\tilde{d}_i\boldsymbol{e}_k^*X_i\mathring{\boldsymbol{g}}_i\frac{\partial\tau_{i3}}{\partial g_{ik}} = O_\prec(\Psi^2\Pi^2)\,, \tag{6.27}$$

$$\frac{1}{N^2}\sum_{i=1}^N\sum_k^{(i)}\tilde{d}_i\boldsymbol{e}_k^*X_i\mathring{\boldsymbol{g}}_i\Big(\frac{1}{N}\sum_{j=1}^N\frac{\partial(d_jQ_j)}{\partial g_{ik}}\Big) = O_\prec(\Psi^2\Pi^2)\,, \tag{6.28}$$

$$\frac{1}{N^2}\sum_{i=1}^N\sum_k^{(i)}\tilde{d}_i\boldsymbol{e}_k^*X_i\mathring{\boldsymbol{g}}_i\Big(\frac{1}{N}\sum_{j=1}^N\frac{\partial(\overline{d_jQ_j})}{\partial g_{ik}}\Big) = O_\prec(\Psi^2\Pi^2)\,. \tag{6.29}$$

The above three bounds follows from the last estimate in (5.56) and the chain rule. Hence, we conclude the proof of (6.23) with $j=1$.

The proof of (6.23) for $j=2$ is similar to $j=1$. Recall the definition of ε_{i2} from (5.52). Using (5.15), (5.16), the first bound in (5.17), and also the bounds in (5.30), we have

$$\varepsilon_{i2} = \big(\|\boldsymbol{g}_i\|^2-1\big)G_{ii}\mathrm{tr}\,\widetilde{B}G + O_\prec\Big(\frac{\Psi}{\sqrt{N}}\Big) = \big(\mathring{\boldsymbol{g}}_i^*\mathring{\boldsymbol{g}}_i - 1\big)\frac{\mathrm{tr}\,\widetilde{B}G}{a_i-\omega_B^c} + O_\prec(\hat{\Pi}^2)\,,$$

which possesses a very similar structure as (6.24). The remaining proof is nearly the same as the case for ε_1; it suffices to replace $\mathring{\boldsymbol{g}}_i^*\widetilde{B}^{\langle i\rangle}\mathring{\boldsymbol{g}}_i$ by $\mathring{\boldsymbol{g}}_i^*\mathring{\boldsymbol{g}}_i$ throughout the proof. We thus omit the details. Hence, we conclude the proof for Lemma 6.3. □

7. Optimal fluctuation averaging

In this section, we establish the optimal fluctuation averaging estimate for a special linear combinations of the Q_i's and their analogues, the $\mathcal{Q}_i$'s (see (7.7) below), under assumption (5.13).

Recall the definitions of the approximate subordination functions ω_A^c and ω_B^c in (5.2). We denote

$$\Lambda_A := \omega_A^c - \omega_A\,, \qquad \Lambda_B := \omega_B^c - \omega_B\,, \qquad \Lambda := |\Lambda_A| + |\Lambda_B|\,. \tag{7.1}$$

Recall $\mathcal{S}_{AB}$, $\mathcal{T}_A$ and $\mathcal{T}_B$ defined in (3.1). For brevity, in the sequel, we use the shorthand notation

$$\mathcal{S} \equiv \mathcal{S}_{AB}.$$

Proposition 7.1. *Fix a $z = E + \mathrm{i}\eta \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$. Suppose that the assumptions of Proposition 5.1 hold. Suppose that $\Lambda(z) \prec \hat{\Lambda}(z)$, for some deterministic and positive function $\hat{\Lambda}(z) \prec N^{-\frac{\gamma}{4}}$, then*

$$\Big|\mathcal{S}\Lambda_\iota + \mathcal{T}_\iota\Lambda_\iota^2 + O(\Lambda_\iota^3)\Big| \prec \frac{\sqrt{(\operatorname{Im} m_{\mu_A\boxplus\mu_B} + \hat{\Lambda})(|\mathcal{S}| + \hat{\Lambda})}}{N\eta} + \frac{1}{(N\eta)^2}, \qquad \iota = A, B\,. \tag{7.2}$$

Before commencing the proof of Proposition 7.1, we first claim that the control parameter $\hat{\Pi}$ in Proposition 6.1 can be chosen as the square root of the right side of (7.2) as long as $\Lambda \prec \hat{\Lambda}$, *i.e.*,

$$\hat{\Pi} := \left(\frac{\sqrt{(\operatorname{Im} m_{\mu_A\boxplus\mu_B} + \hat{\Lambda})(|\mathcal{S}| + \hat{\Lambda})}}{N\eta} + \frac{1}{(N\eta)^2}\right)^{\frac{1}{2}}. \tag{7.3}$$

Indeed, observe that when $\Lambda \prec \hat{\Lambda} \prec N^{-\frac{\gamma}{4}}$, we obtain from the second line of (2.11) that

$$\begin{aligned}
|m_H - m_{\mu_A\boxplus\mu_B}| &= |m_H m_{\mu_A\boxplus\mu_B}|\Big|\frac{1}{m_H(z)} - \frac{1}{m_{\mu_A\boxplus\mu_B}(z)}\Big| \\
&\le C|m_H m_{\mu_A\boxplus\mu_B}|\,\Lambda \\
&\le C|m_H - m_{\mu_A\boxplus\mu_B}|\,\Lambda + C|m_{\mu_A\boxplus\mu_B}|^2\,\Lambda\,,
\end{aligned} \tag{7.4}$$

which together with the fact $|m_{\mu_A\boxplus\mu_B}| \le C$ implies

$$|m_H - m_{\mu_A\boxplus\mu_B}| \prec \Lambda \prec \hat{\Lambda}\,. \tag{7.5}$$

Therefore, recalling (4.9), we have

$$\Pi^2 \prec \frac{\operatorname{Im} m_{\mu_A \boxplus \mu_B} + \hat{\Lambda}}{N\eta} \prec \frac{\sqrt{(\operatorname{Im} m_{\mu_A \boxplus \mu_B} + \hat{\Lambda})(|\mathcal{S}| + \hat{\Lambda})}}{N\eta} \prec \Psi^2,$$

where in the last two steps, we used that $\operatorname{Im} m_{\mu_A \boxplus \mu_B} \lesssim |\mathcal{S}| \prec 1$; (3.4) and (3.5). In addition, from (3.4) and (3.5), we also have $\operatorname{Im} m_{\mu_A \boxplus \mu_B} |\mathcal{S}| \gtrsim \eta$. Thus we also have

$$\frac{1}{N\sqrt{\eta}} \prec \frac{\sqrt{(\operatorname{Im} m_{\mu_A \boxplus \mu_B} + \hat{\Lambda})(|\mathcal{S}| + \hat{\Lambda})}}{N\eta}.$$

From the definition of Π in (4.9), we note that $\Pi \prec \sqrt{\frac{\operatorname{Im} m_{\mu_A \boxplus \mu_B} + \hat{\Lambda}}{N\eta}}$ when $\Lambda \prec \hat{\Lambda}$. Hence, up to a $\frac{1}{N\eta}$ term, $\hat{\Pi}$ defined in (7.3) is a deterministic bound of Π inside the spectrum but it can be much larger than Π in the outside regime where $\mathcal{S} \gg \operatorname{Im} m_{\mu_A \boxplus \mu_B}$ (*cf.*, (3.4) and (3.5)).

With the above notation, we can rewrite (7.2) as

$$\left| \mathcal{S}\Lambda_\iota + \mathcal{T}_\iota \Lambda_\iota^2 + O(\Lambda_\iota^3) \right| \prec \hat{\Pi}^2, \qquad \iota = A, B. \tag{7.6}$$

Recall the definition of Q_i from (4.11). We also introduce their analogues

$$\mathcal{Q}_i \equiv \mathcal{Q}_i(z) := (\widetilde{A}\mathcal{G})_{ii} \operatorname{tr} \mathcal{G} - \mathcal{G}_{ii} \operatorname{tr} \widetilde{A}\mathcal{G}\,, \qquad i \in [\![1, N]\!]\,, \tag{7.7}$$

with $\widetilde{A}$ and $\mathcal{G}$ given in (5.3). To prove Proposition 7.1, we need an optimal fluctuation averaging for a very special combination of the Q_i's and the $\mathcal{Q}_i$'s. To this end, we define the functions $\Phi_1, \Phi_2 : (\mathbb{C}^+)^3 \longrightarrow \mathbb{C}$,

$$\Phi_1(\omega_1, \omega_2, z) := F_A(\omega_2) - \omega_1 - \omega_2 + z\,, \qquad \Phi_2(\omega_1, \omega_2, z) := F_B(\omega_1) - \omega_1 - \omega_2 + z\,, \tag{7.8}$$

where $F_A(\,\cdot\,) \equiv F_{\mu_A}(\,\cdot\,)$ and $F_B(\,\cdot\,) \equiv F_{\mu_B}(\,\cdot\,)$ denote the negative reciprocal Stieltjes transforms of μ_A and μ_B. From the subordination equation (2.11), we have $\Phi_1(\omega_A, \omega_B, z) = \Phi_2(\omega_A, \omega_B, z) = 0$, with $\omega_A \equiv \omega_A(z)$ and $\omega_B \equiv \omega_B(z)$. For brevity, we use the shorthand notations

$$\Phi_1^c := \Phi_1(\omega_A^c, \omega_B^c, z)\,, \qquad \Phi_2^c := \Phi_2(\omega_A^c, \omega_B^c, z)\,. \tag{7.9}$$

Further, we define the quantities

$$\mathcal{Z}_1 := \Phi_1^c + (F_A'(\omega_B) - 1)\Phi_2^c\,, \qquad \mathcal{Z}_2 := \Phi_2^c + (F_B'(\omega_A) - 1)\Phi_1^c\,. \tag{7.10}$$

We are going to show that $\mathcal{Z}_1$ and $\mathcal{Z}_2$ are actually certain linear combinations of the Q_i's and the $\mathcal{Q}_i$'s. We start with the identities

$$\Phi_1^c = -\frac{F_A(\omega_B^c)}{(m_H(z))^2}\frac{1}{N}\sum_{i=1}^N \frac{1}{a_i-\omega_B^c}Q_i\,, \qquad \Phi_2^c = -\frac{F_B(\omega_A^c)}{(m_H(z))^2}\frac{1}{N}\sum_{i=1}^N \frac{1}{b_i-\omega_A^c}\mathcal{Q}_i\,, \tag{7.11}$$

which can be derived by combining (5.2), (5.4) and (5.59). For all $i \in [\![1, N]\!]$, we set

$$\mathfrak{d}_{i,1} := -\frac{F_A(\omega_B^c)}{(m_H(z))^2}\frac{1}{a_i-\omega_B^c}\,, \qquad \mathfrak{d}_{i,2} := -(F_A'(\omega_B)-1)\frac{F_B(\omega_A^c)}{(m_H(z))^2}\frac{1}{b_i-\omega_A^c}\,. \tag{7.12}$$

According to the definition in (7.10), (7.11), and also (7.12), we can write

$$\mathcal{Z}_1 = \frac{1}{N}\sum_{i=1}^N \mathfrak{d}_{i,1}Q_i + \frac{1}{N}\sum_{i=1}^N \mathfrak{d}_{i,2}\mathcal{Q}_i\,, \tag{7.13}$$

and $\mathcal{Z}_2$ can be represented in a similar way.

Now, we choose $d_i = \mathfrak{d}_{i,1}, i \in [\![1, N]\!]$, in Proposition 6.1. Observe that $\mathfrak{d}_{i,1}$ can be regarded as a smooth function of $\operatorname{tr}\widetilde{B}G = 1 - \operatorname{tr}(A-z)G$ and $m_H(z) = \operatorname{tr} G$, according to the definition in (7.12) and that of ω_B^c in (5.2). Then, using the chain rule and the estimates of the tracial quantities in (5.20), one can check that the first equation in assumption (6.2) is satisfied for the choice $d_i = \mathfrak{d}_{i,1}, i \in [\![1, N]\!]$, by using (5.56). The second equation can be checked analogously. Hence, applying Proposition 6.1, we get

$$|\Phi_1^c| \prec \Psi\hat{\Pi}\,, \qquad |\Phi_2^c| \prec \Psi\hat{\Pi}\,, \tag{7.14}$$

where $\hat{\Pi}$ is chosen as in (7.3).

The main technical task in this section is to establish the following estimates for $\mathcal{Z}_1$ and $\mathcal{Z}_2$, where the previous order $\Psi\hat{\Pi}$ bounds from (6.3) are strengthened.

Proposition 7.2. *Fix $z \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$. Suppose that the assumptions of Proposition 5.1 hold and that $\Lambda(z) \prec \hat{\Lambda}(z)$ for some deterministic and positive function $\hat{\Lambda}(z) \le N^{-\frac{\gamma}{4}}$. Choose $\hat{\Pi}(z)$ as (7.3). Then,*

$$|\mathcal{Z}_1| \prec \hat{\Pi}^2\,, \qquad |\mathcal{Z}_2| \prec \hat{\Pi}^2\,. \tag{7.15}$$

We postpone the proof of Proposition 7.2 and first prove Proposition 7.1 with the aid of Proposition 7.2.

Proof of Proposition 7.1. By assumption, we see that $|\Lambda_A|, |\Lambda_B| \prec N^{-\frac{\gamma}{4}}$. First of all, expanding Φ_1^c and Φ_2^c around (ω_A, ω_B) and using the subordination equations $\Phi_1(\omega_A, \omega_B, z) = \Phi_2(\omega_A, \omega_B, z) = 0$, we get

$$\Phi_1^c = -\Lambda_A + (F_A'(\omega_B) - 1)\Lambda_B + \frac{1}{2}F_A''(\omega_B)\Lambda_B^2 + O(\Lambda_B^3)\,,$$
$$\Phi_2^c = -\Lambda_B + (F_B'(\omega_A) - 1)\Lambda_A + \frac{1}{2}F_B''(\omega_A)\Lambda_A^2 + O(\Lambda_A^3)\,. \tag{7.16}$$

We rewrite the second equation in (7.16) as

$$\Lambda_B = -\Phi_2^c + (F_B'(\omega_A) - 1)\Lambda_A + \frac{1}{2}F_B''(\omega_A)\Lambda_A^2 + O(\Lambda_A^3)\,. \tag{7.17}$$

Substituting (7.17) into the first equation in (7.16) yields

$$\Phi_1^c = -(F_A'(\omega_B) - 1)\Phi_2^c + \mathcal{S}\Lambda_A + \mathcal{T}_A\Lambda_A^2 + O((\Phi_2^c)^2) + O(\Phi_2^c\Lambda_A) + O(\Lambda_A^3)\,,$$

where $\mathcal{T}_A$ is defined in (3.1). In light of the definition in (7.10), we have

$$\mathcal{Z}_1 = \mathcal{S}\Lambda_A + \mathcal{T}_A\Lambda_A^2 + O((\Phi_2^c)^2) + O(\Phi_2^c\Lambda_A) + O(\Lambda_A^3)\,. \tag{7.18}$$

Combination of (7.14), (7.15) with (7.18) leads to

$$\left|\mathcal{S}\Lambda_A + \mathcal{T}_A\Lambda_A^2 + O(\Lambda_A^3)\right| \prec \hat{\Pi}^2 + \Psi\hat{\Pi}\hat{\Lambda}\,. \tag{7.19}$$

The second term on the right hand side of (7.19) can be absorbed into the first term, in light of the fact that $\Psi\hat{\Lambda} \prec \hat{\Pi}$ (*cf.*, (7.3)). Hence, we have

$$\left|\mathcal{S}\Lambda_A + \mathcal{T}_A\Lambda_A^2 + O(\Lambda_A^3)\right| \prec \hat{\Pi}^2\,. \tag{7.20}$$

Analogously, we also have

$$\left|\mathcal{S}\Lambda_B + \mathcal{T}_B\Lambda_B^2 + O(\Lambda_B^3)\right| \prec \hat{\Pi}^2\,. \tag{7.21}$$

This completes the proof of Proposition 7.1. $\square$

It remains to prove Proposition 7.2. We state the proof for $\mathcal{Z}_1$, $\mathcal{Z}_2$ is handled similarly. We set

$$\mathfrak{l}^{(k,l)} \equiv \mathfrak{l}^{(k,l)}(z) := \mathcal{Z}_1^k\overline{\mathcal{Z}_1^l}\,, \qquad\qquad k, l \in \mathbb{N}\,.$$

We can now prove a stronger estimate one $\mathbb{E}[\mathfrak{l}^{(p,p)}]$ than the estimate obtained from Lemma 6.2 by improving the error terms from $O_\prec(\Psi\hat{\Pi})$ to $O_\prec(\hat{\Pi}^2)$.

Lemma 7.3. *Fix a $z \in \mathcal{D}_\tau(\eta_\mathrm{m}, \eta_\mathrm{M})$. Suppose that the assumptions of Proposition 7.2 hold. For any fixed integer $p \geq 1$, we have*

$$\mathbb{E}\big[\mathfrak{l}^{(p,p)}(z)\big] = \mathbb{E}\big[O_\prec(\hat{\Pi}^2)\mathfrak{l}^{(p-1,p)}(z)\big] + \mathbb{E}\big[O_\prec(\hat{\Pi}^4)\mathfrak{l}^{(p-2,p)}(z)\big] + \mathbb{E}\big[O_\prec(\hat{\Pi}^4)\mathfrak{l}^{(p-1,p-1)}(z)\big]\,.$$

Now, with Lemma 7.3, we can prove Proposition 7.2.

Proof of Proposition 7.2. Similarly to the proof of (5.14) from Lemma 5.2, with Lemma 7.3, we can get (7.15) by applying Young's and Markov's inequalities. This completes the proof of Proposition 7.2. □

In the sequel, we prove Lemma 7.3.

Proof of Lemma 7.3. Recall the definition of $\mathcal{Z}_1$ in (7.13). We can write

$$\mathbb{E}\left[\mathfrak{l}^{(p,p)}\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\mathfrak{d}_{i,1}Q_i\mathfrak{l}^{(p-1,p)}\right] + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\mathfrak{d}_{i,2}\mathcal{Q}_i\mathfrak{l}^{(p-1,p)}\right].$$

We only state the estimate for the first term on the right hand side above. The second term can be estimated in a similar way. By (6.11), we can write

$$\frac{1}{N}\sum_{i=1}^{N}\mathfrak{d}_{i,1}Q_i = \frac{1}{N}\sum_{i=1}^{N}(\widetilde{B}G)_{ii}\mathrm{tr}\, G\tau_{i1},$$

where we chose $d_i = \mathfrak{d}_{i,1}, i \in [\![1, N]\!]$, in the definition of τ_{i1} in (6.12).

Then, analogously to (6.13), we can also write

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\mathfrak{d}_{i,1}Q_i\mathfrak{l}^{(p-1,p)}\right]&\\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[(\widetilde{B}G)_{ii}\mathrm{tr}\, G\tau_{i1}\mathfrak{l}^{(p-1,p)}\right]\end{aligned}\tag{7.22}$$

with $d_i = \mathfrak{d}_{i,1}, i \in [\![1, N]\!]$. Analogously to (6.5), we can show

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\mathfrak{d}_{i,1}Q_i\mathfrak{l}^{(p-1,p)}\right]\\ &\quad= \mathbb{E}\left[O_\prec(\hat{\Pi}^2)\mathfrak{l}^{(p-1,p)}\right] + \mathbb{E}\left[O_\prec(\Psi^2\hat{\Pi}^2)\mathfrak{l}^{(p-2,p)}\right] + \mathbb{E}\left[O_\prec(\Psi^2\hat{\Pi}^2)\mathfrak{l}^{(p-1,p-1)}\right],\end{aligned}$$

where the last two terms come from the estimates of the analogues of the last two terms of (6.16), the third and fourth terms in the right side of (6.20), and also the terms in (6.28) and (6.29), but with $\frac{1}{N}\sum_{j=1}^{N} d_jQ_j$ replaced by $\mathcal{Z}_1$. It suffices to improve the estimates of these terms. All these terms contain a derivative $\frac{\partial \mathcal{Z}_1}{\partial g_{ik}}$ or $\frac{\partial \overline{\mathcal{Z}}_1}{\partial g_{ik}}$, which is smaller than the derivative of an arbitrary linear combination $\partial(\frac{1}{N}\sum_i d_iQ_i)/\partial g_{ik}$ or $\partial(\frac{1}{N}\sum_i d_i\mathcal{Q}_i)/\partial g_{ik}$, due to the special choice of $\mathfrak{d}_{i,1}$'s and $\mathfrak{d}_{i,2}$'s. Specifically, we shall show the following lemma, which contains the estimates of all necessary terms.

Lemma 7.4. *Fix a $z \in \mathcal{D}_\tau(\eta_\mathrm{m}, \eta_\mathrm{M})$. Suppose that the assumptions of Proposition 5.1 hold. Let $\tilde{d}_1, \ldots, \tilde{d}_N \in \mathbb{C}$ be (possibly random) numbers with $\max_i |\tilde{d}_i| \prec 1$. Let $X_i = I$ or $\widetilde{B}^{\langle i\rangle}$. Then we have*

$$\frac{1}{N^2}\sum_{i=1}^N \sum_k^{(i)} \tilde{d}_i \boldsymbol{e}_k^* X_i G \boldsymbol{e}_i \frac{\partial \mathcal{Z}_1}{\partial g_{ik}} = O_\prec(\hat{\Pi}^4)\,, \qquad \frac{1}{N^2}\sum_{i=1}^N \sum_k^{(i)} \tilde{d}_i \boldsymbol{e}_k^* X_i G \boldsymbol{e}_i \frac{\partial \overline{\mathcal{Z}_1}}{\partial g_{ik}} = O_\prec(\hat{\Pi}^4)\,,$$
$$\frac{1}{N^2}\sum_{i=1}^N \sum_k^{(i)} \tilde{d}_i \boldsymbol{e}_k^* X_i \mathring{\boldsymbol{g}}_i \frac{\partial \mathcal{Z}_1}{\partial g_{ik}} = O_\prec(\hat{\Pi}^4)\,, \qquad \frac{1}{N^2}\sum_{i=1}^N \sum_k^{(i)} \tilde{d}_i \boldsymbol{e}_k^* X_i \mathring{\boldsymbol{g}}_i \frac{\partial \overline{\mathcal{Z}_1}}{\partial g_{ik}} = O_\prec(\hat{\Pi}^4)\,. \tag{7.23}$$

Proof of Lemma 7.4. We give the proof for the first estimate in (7.23). The third one is analogous, and the other two are just their complex conjugates. From the definitions in (7.9) and (7.10), we get

$$\begin{aligned}\frac{\partial \mathcal{Z}_1}{\partial g_{ik}} &= \frac{\partial \Phi_1^c}{\partial g_{ik}} + (F_A'(\omega_B) - 1)\frac{\partial \Phi_2^c}{\partial g_{ik}} \\ &= \Big(\big(F_A'(\omega_B) - 1\big)\big(F_B'(\omega_A^c) - 1\big) - 1\Big)\frac{\partial \omega_A^c}{\partial g_{ik}} + \big(F_A'(\omega_B^c) - F_A'(\omega_B)\big)\frac{\partial \omega_B^c}{\partial g_{ik}}.\end{aligned}$$

Note that by the regularity of F_A and F_B, we have

$$\big(F_A'(\omega_B) - 1\big)\big(F_B'(\omega_A^c) - 1\big) - 1 = \mathcal{S} + O(|\Lambda_A|)\,, \qquad F_A'(\omega_B^c) - F_A'(\omega_B) = O(|\Lambda_B|)\,.$$

The smallness of these coefficients carry the gain. According to the definition of $\hat{\Pi}$ in (7.3), we see that

$$(|\mathcal{S}| + \Lambda)\Psi^2\Pi^2 \le \hat{\Pi}^4$$

if $\Lambda \le \hat{\Lambda}$. Hence, for the first estimate in (7.23), it suffices to show that

$$\frac{1}{N^2}\sum_{i=1}^N \sum_k^{(i)} \tilde{d}_i \boldsymbol{e}_k^* X_i G \boldsymbol{e}_i \frac{\partial \omega_\iota^c}{\partial g_{ik}} = O_\prec(\Psi^2\Pi^2)\,, \qquad \iota = A, B\,. \tag{7.24}$$

This follows from (6.22), the fact that ω_B^c is a tracial quantity, and the chain rule. The other terms in (7.23) can be estimated similarly. This concludes the proof of Lemma 7.4. □

With the aid of Lemma 7.4, we can conclude the proof of Lemma 7.3. □

8. Weak local law

In Sections 5, 6 and 7, we established the subordination property for the Green function entries and the rough and optimal fluctuation averaging for the linear combinations of them, but all for a fixed $z \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$, under the a priori input (5.13). In this section, based on some cutoff versions of the conclusions in Sections 5 and 6 (*cf.*, (8.17), (8.29)), we will establish a weak local law, uniformly in $z \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$, by using a continuity argument. The weak local law will guarantee that the input in (5.13) hold uniformly true on $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$, and thus the conclusions in Sections 5, 6 and 7 are also uniformly true on $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$.

Our main result in this section is the following weak local law for the quantities P_i, K_i, T_i, $\Lambda_{\mathrm{d}i}^c(z)$, Υ, Λ_{d}, Λ, defined in (4.14), (4.15), (4.10), (5.7), (5.10), (5.6), (7.1), respectively.

Theorem 8.1 *(Weak local law at the regular edge). Suppose that Assumptions 2.1 and 2.2 hold. Let $\tau > 0$ be a sufficiently small constant and fix any (small) constants $\gamma > 0$. Then, for all $i \in [\![1, N]\!]$, we have*

$$|P_i(z)| \prec \Psi(z)\,, \quad |K_i(z)| \prec \Psi(z)\,, \quad \Lambda_{\mathrm{d}i}^c(z) \prec \Psi(z)\,, \quad |T_i| \prec \Psi(z)\,, \quad |\Upsilon(z)| \prec \Psi(z)\,, \tag{8.1}$$

uniformly in $z \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$. In addition, we have

$$\Lambda_{\mathrm{d}}(z) \prec \frac{1}{(N\eta)^{\frac{1}{3}}}\,, \qquad \Lambda(z) \prec \frac{1}{(N\eta)^{\frac{1}{3}}}\,, \tag{8.2}$$

uniformly in $z \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$.

The same statements hold for the analogous quantities with tildes (see their definitions around (5.8)), i.e. *if we switch the roles of A and B, and also the roles of U and U^*.*

In order to prove Theorem 8.1, we first need the following lemma. Recall from (7.1) the definitions $\Lambda_\iota(z) = \omega_\iota(z) - \omega_\iota^c(z)$, $\iota = A, B$. Further recall the definitions of $\mathcal{T}_\iota$, $\iota = A, B$, and $\mathcal{S}_{AB}$ from (3.1) and that we abbreviate $\mathcal{S} = \mathcal{S}_{AB}$.

Lemma 8.2. *Fix $z \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$. Let $\varepsilon \in (0, \frac{\gamma}{100})$. Let $\hat{\Lambda} \equiv \hat{\Lambda}(z)$ be some deterministic control parameter satisfying $\frac{N^{3\varepsilon}}{(N\eta)^{\frac{1}{3}}} \leq \hat{\Lambda}(z) \leq N^{-\frac{\gamma}{4}}$. Suppose that $\Lambda \leq \hat{\Lambda}$ and*

$$\Big|\mathcal{S}\Lambda_\iota + \mathcal{T}_\iota\Lambda_\iota^2 + O(\Lambda_\iota^3)\Big| \leq N^{\varepsilon}\frac{|\mathcal{S}| + \hat{\Lambda}}{(N\eta)^{\frac{1}{3}}}, \qquad \iota = A, B \tag{8.3}$$

hold on some event $\widetilde{\Omega}(z)$. Then we have, for N sufficiently large.
(i): If $\sqrt{\kappa + \eta} > N^{-\varepsilon}\hat{\Lambda}$, there is a sufficiently large constant $K_0 > 0$ independent of z, such that

$$\mathbf{1}\Big(\Lambda \leq \frac{|\mathcal{S}|}{K_0}\Big)|\Lambda_A| \leq N^{-2\varepsilon}\hat{\Lambda}\,, \qquad \mathbf{1}\Big(\Lambda \leq \frac{|\mathcal{S}|}{K_0}\Big)|\Lambda_B| \leq N^{-2\varepsilon}\hat{\Lambda} \qquad \text{on} \quad \widetilde{\Omega}(z)\,, \tag{8.4}$$

where $\mathbf{1}$ *denotes the indicator function.*
(ii): If $\sqrt{\kappa+\eta} \le N^{-\varepsilon}\hat{\Lambda}$, *we have*

$$|\Lambda_A| \le N^{-\varepsilon}\hat{\Lambda}\,, \qquad\qquad |\Lambda_B| \le N^{-\varepsilon}\hat{\Lambda} \qquad \textit{on} \quad \widetilde{\Omega}(z). \tag{8.5}$$

Proof of Lemma 8.2. We first recall (3.5). Then, from the assumptions $|\Lambda_\iota| \le \hat{\Lambda} \le N^{-\frac{\gamma}{4}}$ and (8.3), we have on the event $\widetilde{\Omega}(z)$ that

$$\mathcal{S}\Lambda_\iota + \mathcal{T}_\iota\Lambda_\iota^2 = O\Big(N^{\varepsilon}\frac{|\mathcal{S}|+\hat{\Lambda}}{(N\eta)^{\frac13}} + N^{-\frac{\gamma}{4}}\hat{\Lambda}^2\Big), \qquad \iota = A, B. \tag{8.6}$$

If $\sqrt{\kappa+\eta} > N^{-\varepsilon}\hat{\Lambda}$, we have for $\iota = A, B$, and sufficiently large constant $K_0 > 0$,

$$\mathbf{1}\Big(\Lambda \le \frac{|\mathcal{S}|}{K_0}\Big)|\Lambda_\iota| \le C|\mathcal{S}|^{-1}\Big(N^{\varepsilon}\frac{|\mathcal{S}|+\hat{\Lambda}}{(N\eta)^{\frac13}} + N^{-\frac{\gamma}{4}}\hat{\Lambda}^2\Big) \le C\frac{N^{\varepsilon}}{(N\eta)^{\frac13}} + CN^{\varepsilon-\frac{\gamma}{4}}\hat{\Lambda} \le CN^{-2\varepsilon}\hat{\Lambda}\,. \tag{8.7}$$

Here we absorbed the quadratic term on the left side of (8.6) into the linear term and used that $\mathcal{S} \sim \sqrt{\kappa+\eta}$ and $|\mathcal{T}_\iota| \lesssim 1$; see Proposition 3.1. Hence, we proved (i). From (8.7), we also see that if $\sqrt{\kappa+\eta} > N^{-\varepsilon}\hat{\Lambda}$, then

$$\mathbf{1}\Big(\Lambda \le \frac{|\mathcal{S}|}{K_0}\Big)|\Lambda_\iota| \le CN^{-\varepsilon}|\mathcal{S}|\,, \qquad\qquad \iota = A, B. \tag{8.8}$$

Next, we prove (ii). If $\sqrt{\kappa+\eta} \le N^{-\varepsilon}\hat{\Lambda}$, from (3.5) and (3.6), we see that $\mathcal{T}_\iota \sim 1$. Hence, we solve the quadratic equation (8.6) directly, then we get

$$|\Lambda_\iota| \le C|\mathcal{S}| + C\Big(N^{\varepsilon}\frac{|\mathcal{S}|+\hat{\Lambda}}{(N\eta)^{\frac13}} + N^{-\frac{\gamma}{4}}\hat{\Lambda}^2\Big)^{\frac12} \le CN^{-\varepsilon}\hat{\Lambda}\,, \qquad \iota = A, B\,,$$

under the assumption that $\hat{\Lambda} \ge \frac{N^{3\varepsilon}}{(N\eta)^{\frac13}}$. This concludes the proof of Lemma 8.2. $\square$

Recall the definitions of Λ_{d}, $\widetilde{\Lambda}_{\mathrm{d}}$, Λ_T, $\widetilde{\Lambda}_T$ in (5.6). For any $z \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$ and any $\delta, \delta' \in [0,1]$, we define the event

$$\Theta(z,\delta,\delta') := \Big\{\Lambda_{\mathrm{d}}(z) \le \delta,\ \widetilde{\Lambda}_{\mathrm{d}}(z) \le \delta,\ \Lambda(z) \le \delta,\ \Lambda_T(z) \le \delta',\ \widetilde{\Lambda}_T(z) \le \delta'\Big\}. \tag{8.9}$$

We further decompose the domain $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$ into the following two disjoint parts:

$$\begin{aligned} \mathcal{D}_> &:= \Big\{z \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}}) : \sqrt{\kappa+\eta} > \frac{N^{2\varepsilon}}{(N\eta)^{\frac13}}\Big\}, \\ \mathcal{D}_\le &:= \Big\{z \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}}) : \sqrt{\kappa+\eta} \le \frac{N^{2\varepsilon}}{(N\eta)^{\frac13}}\Big\}. \end{aligned} \tag{8.10}$$

For $z \in \mathcal{D}_>$, any $\delta, \delta' \in [0,1]$ and any $\varepsilon' \in [0,1]$, we define the event $\Theta_>(z,\delta,\delta',\varepsilon') \subset \Theta(z,\delta,\delta')$ as

$$\begin{aligned}&\Theta_>(z,\delta,\delta',\varepsilon')\\ &\quad := \Big\{\Lambda_{\mathrm{d}}(z) \le \delta,\ \widetilde{\Lambda}_{\mathrm{d}}(z) \le \delta,\ \Lambda(z) \le \min\{\delta, N^{-\varepsilon'}|\mathcal{S}|\},\ \Lambda_T(z) \le \delta',\ \widetilde{\Lambda}_T(z) \le \delta'\Big\}.\end{aligned} \tag{8.11}$$

Lemma 8.3. *Suppose that the assumptions in Theorem 2.5 hold. For any fixed $z \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$, any $\varepsilon \in (0, \frac{\gamma}{100})$ and any $D > 0$, there exists a positive integer $N_1(D,\varepsilon)$ and an event $\Omega(z) \equiv \Omega(z,D,\varepsilon)$ with*

$$\mathbb{P}(\Omega(z)) \ge 1 - N^{-D}, \qquad \forall N \ge N_1(D,\varepsilon), \tag{8.12}$$

such that the following hold:

(i) If $z \in \mathcal{D}_>$, we have

$$\Theta_>\Big(z, \frac{N^{3\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{3\varepsilon}}{\sqrt{N\eta}}, \frac{\varepsilon}{10}\Big) \cap \Omega(z) \subset \Theta_>\Big(z, \frac{N^{\frac{5}{2}\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{\frac{5}{2}\varepsilon}}{\sqrt{N\eta}}, \frac{\varepsilon}{2}\Big). \tag{8.13}$$

(ii) If $z \in \mathcal{D}_\le$, we have

$$\Theta\Big(z, \frac{N^{3\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{3\varepsilon}}{\sqrt{N\eta}}\Big) \cap \Omega(z) \subset \Theta\Big(z, \frac{N^{\frac{5}{2}\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{\frac{5}{2}\varepsilon}}{\sqrt{N\eta}}\Big). \tag{8.14}$$

Proof of Lemma 8.3. The proof relies on a cutoff version of Lemma 5.2 and Lemma 6.2, where we will introduce some smooth cutoff to $\mathfrak{m}_i$ and $\mathfrak{m}$ to guarantee that the a priori inputs needed for the estimates hold. The same idea has already been used in Lemma 5.5 of [6], but for completeness we repeat the arguments. With these cutoff versions of Lemmas 5.2 and 6.2, the proof of Lemma 8.3 is accomplished in three steps, corresponding to what we did in Sections 5, 6 and 7, respectively.

Step 1: In this step we establish the cutoff version of Lemma 5.2 and use it to prove an estimate of for G_{ii}'s, T_i's and their tilde analogues. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a smooth cutoff function equal to 1 on $[-\mathcal{L}, \mathcal{L}]$ and vanishing on $[-2\mathcal{L}, 2\mathcal{L}]^c$, such that $\sup_{x \in \mathbb{R}} |\varphi'(x)| \le C\mathcal{L}^{-1}$ for some sufficiently large constant $\mathcal{L} > 0$. Let

$$\Gamma_i \equiv \Gamma_i(z) := |G_{ii}|^2 + |\mathcal{G}_{ii}|^2 + |T_i|^2 + |\widetilde{T}_i|^2 + |\operatorname{tr} G|^2 + |\operatorname{tr} \widetilde{B}G|^2 + |\operatorname{tr} \widetilde{B}G\widetilde{B}|^2, \tag{8.15}$$

where we denote by $\widetilde{T}_i$ the analogue of T_i, obtained via switching the roles of A and B and also the roles of U and U^* in the definition of T_i (*cf.*, (4.10)). For a given i, observe that all the a priori inputs we needed in the proof of Lemma 5.2 are the $O_\prec(1)$ bound for the summands on the right side of (8.15). Recall the definitions of $\mathfrak{m}_i^{(k,\ell)}$ and $\mathfrak{n}_i^{(k,\ell)}$ from (5.23), and set

$$\widetilde{\mathfrak{m}}_i^{(k,\ell)} := \mathfrak{m}_i^{(k,\ell)}(\varphi(\Gamma_i))^{k+\ell}, \qquad \widetilde{\mathfrak{n}}_i^{(k,\ell)} := \mathfrak{n}_i^{(k,\ell)}(\varphi(\Gamma_i))^{k+\ell} \tag{8.16}$$

In addition, for any $\varepsilon_1 > 0$, let $\widehat{\Omega}_1(z) \equiv \widehat{\Omega}_1(z, \varepsilon_1)$ be the event that all concentration estimates of the components or quadratic forms of the Gaussian vectors $\boldsymbol{g}_i$'s in the proof of Lemma 5.2 hold with precision N^{ε_1}. For instance, we used the $O_\prec(\frac{1}{\sqrt{N}})$ bound for $\boldsymbol{h}_i^* \widetilde{B}^{\langle i\rangle} \boldsymbol{h}_i$ in (5.30). Now we can bound it more quantitatively by $\frac{N^{\varepsilon_1}}{\sqrt{N}}$ on $\widehat{\Omega}_1(z)$. Due to the Gaussian tail, for any $D_1 > 0$, there exists an $N(D_1, \varepsilon_1)$, such that if $N \geq N(D_1, \varepsilon_1)$, then

$$\mathbb{P}(\widehat{\Omega}_1(z)) \geq 1 - N^{-D_1}.$$

Furthermore, we claim that

$$\mathbb{E}[\widetilde{\mathfrak{m}}_i^{(p,p)}] = \mathbb{E}[\mathfrak{c}_{i1} \widetilde{\mathfrak{m}}_i^{(p-1,p)}] + \mathbb{E}[\mathfrak{c}_{i2} \widetilde{\mathfrak{m}}_i^{(p-2,p)}] + \mathbb{E}[\mathfrak{c}_{i3} \widetilde{\mathfrak{m}}_i^{(p-1,p-1)}], \tag{8.17}$$

with some random variables $\mathfrak{c}_{i1}, \mathfrak{c}_{i2}$ and $\mathfrak{c}_{i3}$, satisfying

$$|\mathfrak{c}_{i1}| \leq C \frac{N^{\varepsilon_1}}{\sqrt{N\eta}}, \qquad |\mathfrak{c}_{i2}| \leq C \frac{N^{2\varepsilon_1}}{N\eta}, \qquad |\mathfrak{c}_{i3}| \leq C \frac{N^{2\varepsilon_1}}{N\eta}, \qquad \text{on} \quad \widehat{\Omega}_1(z),$$

for some positive constant C which may depend on $\mathcal{L}$, *i.e.*, the parameter in the definition of the cutoff φ. Moreover, the $\mathfrak{c}_{ia}$'s also admit the moment bound $\mathbb{E}|\mathfrak{c}_{ia}|^k = O(1)$ for any given $k > 0$. Note that (8.17) is the same as (5.24) but with a cutoff, since $\widehat{\Omega}_1(z)$ holds with high probability. The proof of (8.17) can be done in the same way as the proof of the non-cutoff one in (5.24). Essentially, the only modification we need to accommodate is estimating the additional terms in the integration by parts that are created by introducing $\varphi(\Gamma_i)$ into $\widetilde{\mathfrak{m}}_i$. But it will be clear that these additional terms can be absorbed into the first term on the right side of (8.17). For instance, in the analogue of the step (5.34), apart from replacing $\mathfrak{m}_i$ by $\widetilde{\mathfrak{m}}_i$, we will have an additional term

$$\frac{1}{N} \sum_k^{(i)} \mathbb{E}\Big[\frac{\boldsymbol{e}_k^* \widetilde{B}^{\langle i\rangle} G \boldsymbol{e}_i}{\|\boldsymbol{g}_i\|} \operatorname{tr} G \frac{\partial \varphi(\Gamma_i)}{\partial g_{ik}} \widetilde{\mathfrak{m}}_i^{(p-1,p)} \Big]. \tag{8.18}$$

According to the definition of $\varphi(\Gamma_i)$, the derivative $\frac{\partial \varphi(\Gamma_i)}{\partial g_{ik}}$ is written as a sum of several terms. For instance, one term is

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

Applying a quantitative version of the second estimate in (5.56), one obtains

$$\frac{1}{N} \sum_k^{(i)} \frac{\boldsymbol{e}_k^* \widetilde{B}^{\langle i\rangle} G \boldsymbol{e}_i}{\|\boldsymbol{g}_i\|} \operatorname{tr} G \frac{\partial |G_{ii}|^2}{\partial g_{ik}} = O\Big(\frac{N^{\varepsilon_1}}{\sqrt{N\eta}}\Big), \qquad \text{on} \;\; \{\varphi'(\Gamma_i) \neq 0\} \cap \widehat{\Omega}_1(z).$$

The contribution of the other terms from $\frac{\partial \varphi(\Gamma_i)}{\partial g_{ik}}$ to the term (8.18) can be bounded in the same way. We omit the details. Hence, we have (8.17).

Applying Young's inequality to (8.17), we have

$$\mathbb{E}[\widetilde{\mathfrak{m}}_i^{(p,p)}] \leq C_p N^{2p\varepsilon_1}\Big(\mathbb{E}|\mathfrak{c}_1|^{2p} + \mathbb{E}|\mathfrak{c}_2|^p + \mathbb{E}|\mathfrak{c}_3|^p\Big) \leq C_p N^{2p\varepsilon_1}\Big(\big(\frac{N^{\varepsilon_1}}{\sqrt{N\eta}}\big)^{2p} + N^{-\frac{D_1}{2}}\Big).$$

Further, by Markov's inequality, we have

$$\mathbb{P}\Big(|P_i\varphi(\Gamma_i)| \geq \frac{N^{\frac{\varepsilon}{4}}}{\sqrt{N\eta}}\Big) \leq C_p\Big(\frac{N^{\frac{\varepsilon}{4}}}{\sqrt{N\eta}}\Big)^{-2p} N^{2p\varepsilon_1}\Big(\big(\frac{N^{\varepsilon_1}}{\sqrt{N\eta}}\big)^{2p} + N^{-\frac{D_1}{2}}\Big). \tag{8.19}$$

For the given $\varepsilon > 0$ in Lemma 8.3, we can first choose $\varepsilon_1 = \varepsilon_1(\varepsilon)$ to be smaller than $\frac{\varepsilon}{8}$, and then choose $p = p(\varepsilon, D)$ to be sufficiently large, then we can get

$$C_p\Big(\frac{N^{\frac{\varepsilon}{4}}}{\sqrt{N\eta}}\Big)^{-2p} N^{2p\varepsilon_1}\Big(\frac{N^{\varepsilon_1}}{\sqrt{N\eta}}\Big)^{2p} \leq \frac{1}{10}N^{-D}. \tag{8.20}$$

Then by choosing $D_1 = D_1(\varepsilon, D)$ sufficiently large, we also have

$$C_p\Big(\frac{N^{\frac{\varepsilon}{4}}}{\sqrt{N\eta}}\Big)^{-2p} N^{2p\varepsilon_1} N^{-\frac{D_1}{2}} \leq \frac{1}{10}N^{-D}. \tag{8.21}$$

We can thus denote $\widetilde{N}_1(D,\varepsilon) := N(D_1,\varepsilon_1)$, according to our choice of ε_1, D_1. Then by (8.19)-(8.21), there exists an event $\Omega_1(z)$, such that

$$\mathbb{P}(\Omega_1(z)) \geq 1 - \frac{1}{5}N^{-D}, \qquad N \geq \widetilde{N}_1(D,\varepsilon)$$

and

$$|P_i\varphi(\Gamma_i)| \leq \frac{N^{\frac{\varepsilon}{4}}}{\sqrt{N\eta}}, \qquad \text{on } \ \Omega_1(z). \tag{8.22}$$

Observing that $\varphi(\Gamma_i) = 1$ on $\Theta(z, \frac{N^{3\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{3\varepsilon}}{\sqrt{N\eta}})$, we further get from (8.22) that

$$|P_i| \leq \frac{N^{\frac{\varepsilon}{4}}}{\sqrt{N\eta}}, \qquad \text{on } \ \Theta\Big(z, \frac{N^{3\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{3\varepsilon}}{\sqrt{N\eta}}\Big) \cap \Omega_1(z). \tag{8.23}$$

By working with $\widetilde{\mathfrak{n}}_i$ instead of $\widetilde{\mathfrak{m}}_i$, we can also get

$$|K_i| \leq \frac{N^{\frac{\varepsilon}{4}}}{\sqrt{N\eta}}, \qquad \text{on } \ \Theta\Big(z, \frac{N^{3\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{3\varepsilon}}{\sqrt{N\eta}}\Big) \cap \Omega_1(z)\,, \tag{8.24}$$

where we redefine $\Omega_1(z)$ to include the "good" events for both estimates for P_i and K_i. Similarly to the proof of Proposition 5.1, based on the estimates in (8.23), (8.24) and also their analogues by switching the roles of A and B, and also the roles of U and U^*, we can derive the following quantitative estimates:

$$\Lambda_{\mathrm{d}}^c(z) \le \frac{N^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \quad \widetilde{\Lambda}_{\mathrm{d}}^c(z) \le \frac{N^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \quad \Lambda_T(z) \le \frac{N^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \quad \widetilde{\Lambda}_T(z) \le \frac{N^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \quad |\Upsilon(z)| \le \frac{N^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \tag{8.25}$$

hold on $\Theta(z, \frac{N^{3\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{3\varepsilon}}{\sqrt{N\eta}}) \cap \Omega_1(z)$, where $\Omega_1(z)$ shall be redefined further to include the "good" events for both estimates of the analogues of P_i and K_i.

Step 2: In this step, we derive a cutoff version of Lemma 6.2, but with weaker bounds, and use it to estimate the linear combinations of G_{ii}. Analogously to (8.16), we introduce the variant of (6.4)

$$\widetilde{\mathfrak{m}}^{(k,l)} := \Big(\frac{1}{N}\sum_{i=1}^N d_i Q_i \varphi(\Gamma_i)\varphi(\Gamma)\Big)^k \Big(\frac{1}{N}\sum_{i=1}^N \overline{d_i}\,\overline{Q_i}\varphi(\Gamma_i)\varphi(\Gamma)\Big)^l, \qquad k,l \in \mathbb{N}, \tag{8.26}$$

where Γ_i is defined in (8.15) and Γ is defined as the following

$$\begin{aligned}\Gamma := (c\operatorname{Im} m_{\mu_A \boxplus \mu_B} + \hat{\Lambda})^{-2}\big(|\Lambda_A|^2 + |\Lambda_B|^2\big) + \Big(\frac{N^{5\varepsilon}}{(N\eta)^{\frac{1}{3}}}\Big)^{-2}|\Upsilon|^2 \\ + \Big(\frac{N^{5\varepsilon}}{\sqrt{N\eta}}\Big)^{-1}\frac{1}{N}\sum_{i=1}^N(|T_i|^2 + N^{-1})^{\frac{1}{2}},\end{aligned} \tag{8.27}$$

for some sufficiently small constant $c > 0$. In the rest of the proof, we choose

$$\hat{\Lambda}(z) = \frac{N^{3\varepsilon}}{(N\eta)^{\frac{1}{3}}}. \tag{8.28}$$

The boundedness of the first term in (8.27) is used to control Π by $\hat{\Pi}$, see (7.3) for its definition. The second and the third terms in (8.27) are used to bound the analogue of the term $\frac{1}{N}\sum_i T_i \tau_{i1}\Upsilon$ in (6.18).

Similarly to $\widehat{\Omega}_1(z)$, for any $\varepsilon_1 > 0$, let $\widehat{\Omega}_2(z) \equiv \widehat{\Omega}_2(z, \varepsilon_1)$ be the event that all concentration estimates of the components or quadratic forms of the Gaussian vectors $\boldsymbol{g}_i$'s in the proof of Lemma 6.2 hold with precision N^{ε_1}. Again, due to the Gaussian tails, for any $D_1 > 0$, there exists $\widetilde{N}(D_1, \varepsilon_1)$, such that if $N \ge \widetilde{N}(D_1, \varepsilon_1)$

$$\mathbb{P}(\widehat{\Omega}_2(z)) \ge 1 - N^{-D_1}.$$

Analogously to (8.17), we now claim that

$$\mathbb{E}[\widetilde{\mathfrak{m}}^{(p,p)}] = \mathbb{E}[\mathfrak{c}_1\widetilde{\mathfrak{m}}^{(p-1,p)}] + \mathbb{E}[\mathfrak{c}_2\widetilde{\mathfrak{m}}^{(p-2,p)}] + \mathbb{E}[\mathfrak{c}_3\widetilde{\mathfrak{m}}^{(p-1,p-1)}], \tag{8.29}$$

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

$$|\mathfrak{c}_1| \le C\hat{\Pi}\,, \qquad |\mathfrak{c}_2| \le C\hat{\Pi}^2\,, \qquad |\mathfrak{c}_3| \le C\hat{\Pi}^2\,, \qquad \text{on} \quad \widehat{\Omega}_2(z)\,, \tag{8.30}$$

for some positive constant C. Moreover, the $\mathfrak{c}_i$'s also admit the moment bound $\mathbb{E}|\mathfrak{c}_i|^k = O(1)$, for any given $k > 0$. Note that (8.29) is similar to (6.5), but with a weaker bounds for $\mathfrak{c}_i$'s. The weakness of the bounds is partially due to the weak a priori input in the cutoffs $\varphi(\Gamma_i)$ and $\varphi(\Gamma)$, and also partially due to the additional terms involving the derivatives of the cutoffs which are generated by the integration by parts. In Appendix C, we show more details on how to slightly modify the proof of (6.5) to get (8.29). Similarly to (8.22), we can show from (8.29) that there exists an event $\Omega_2(z)$, such that

$$\mathbb{P}(\Omega_2(z)) \ge 1 - \frac{1}{5}N^{-D}\,, \qquad N \ge \widetilde{N}_2(D,\varepsilon)\,,$$

for some sufficiently large constant $\widetilde{N}_2(D,\varepsilon) > 0$, and

$$\Big|\frac{1}{N}\sum_{i=1}^N d_i Q_i \varphi(\Gamma_i)\varphi(\Gamma)\Big| \le N^{\frac{\varepsilon}{4}}\hat{\Pi}, \qquad \text{on} \quad \Omega_2(z)\,.$$

Note that $\varphi(\Gamma_i) = \varphi(\Gamma) = 1$ for all i on $\Theta(z, \frac{N^{3\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{3\varepsilon}}{\sqrt{N\eta}})$. Hence, we have

$$\Big|\frac{1}{N}\sum_{i=1}^N d_i Q_i\Big| \le N^{\frac{\varepsilon}{3}}\hat{\Pi}\,, \qquad \text{on} \quad \Theta\Big(z, \frac{N^{3\varepsilon}}{(N\eta)^{\frac{1}{3}}}\,, \frac{N^{3\varepsilon}}{\sqrt{N\eta}}\Big) \cap \Omega_2(z)\,.$$

Step 3: In the last step, we perform an estimate of Λ_A and Λ_B, by choosing $d_i = \mathfrak{d}_{i1}$ in (7.12) and also considering the analogues of $\frac{1}{N}\sum_{i=1}^N \mathfrak{d}_{i1}Q_i$. Repeating the argument of the proof of Proposition 7.2 but with the cutoff versions, where the error terms due to the cutoff are estimated analogously as in Appendix C, we can then finally show that there exists an event $\Omega_3(z)$, such that

$$\mathbb{P}(\Omega_3(z)) \ge 1 - \frac{4}{5}N^{-D}, \qquad N \ge \widetilde{N}_3(D,\varepsilon)\,,$$

for some sufficiently large constant $\widetilde{N}_3(D,\varepsilon) > 0$, and

$$\big|\mathcal{Z}_\iota\big| \le N^{\frac{\varepsilon}{3}}\hat{\Pi} \le N^{\frac{\varepsilon}{2}}\frac{|\mathcal{S}| + \hat{\Lambda}}{(N\eta)^{\frac{1}{3}}}, \quad \iota = 1,2 \qquad \text{on} \quad \Theta\Big(z, \frac{N^{3\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{3\varepsilon}}{\sqrt{N\eta}}\Big) \cap \Omega_3(z),$$

where $\mathcal{Z}_\iota$'s are defined in (7.10). Here in the second step above we used the fact $\operatorname{Im} m_{\mu_A \boxplus \mu_B} \lesssim |\mathcal{S}|$ (*cf.*, (3.4), (3.5)), and thus

$$\hat{\Pi} \leq C\Big(\frac{|\mathcal{S}|+\hat{\Lambda}}{N\eta}\Big)^{\frac{1}{2}} \leq C\Big(\frac{|\mathcal{S}|+\hat{\Lambda}}{(N\eta)^{\frac{1}{3}}}+\frac{1}{(N\eta)^{\frac{2}{3}}}\Big) \leq C'\frac{|\mathcal{S}|+\hat{\Lambda}}{(N\eta)^{\frac{1}{3}}}$$

under the choice of $\hat{\Lambda}$ in (8.28). Similarly to the proof of (7.2), we can then get the following weaker but quantitative version of (7.2)

$$\Big|\mathcal{S}\Lambda_\iota + \mathcal{T}_\iota\Lambda_\iota^2 + O(\Lambda_\iota^3)\Big| \leq N^{\varepsilon}\frac{|\mathcal{S}|+\hat{\Lambda}}{(N\eta)^{\frac{1}{3}}}, \quad \iota = 1,2 \qquad \text{on} \quad \Theta(z, \frac{N^{3\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{3\varepsilon}}{\sqrt{N\eta}}) \cap \Omega_3(z)\,.$$

Then, applying Lemma 8.2, we have (8.4) and (8.5) with the choice $\widetilde{\Omega}(z) = \Theta(z, \frac{N^{3\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{3\varepsilon}}{\sqrt{N\eta}}) \cap \Omega_3(z)$. In addition, from the conclusion of (8.4) and recalling that $|\mathcal{S}| \sim \sqrt{\kappa+\eta}$, we note that if $\sqrt{\kappa+\eta} > N^{-\varepsilon}\hat{\Lambda}$, then

$$\mathbf{1}\Big(\Lambda \leq \frac{|\mathcal{S}|}{K_0}\Big)|\Lambda_\iota| \leq N^{-\varepsilon}|\mathcal{S}|, \qquad \iota = A, B\,, \tag{8.31}$$

with K_0 chosen in Lemma 8.2.

Therefore, with the choice in (8.28), by (8.4), (8.5) and (8.31), we have

$$\Lambda \leq \min\Big\{\frac{N^{\frac{5}{2}\varepsilon}}{(N\eta)^{\frac{1}{3}}}, N^{-\varepsilon}|\mathcal{S}|\Big\}, \qquad \text{on} \quad \Theta_>\Big(z, \frac{N^{3\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{3\varepsilon}}{\sqrt{N\eta}}, \frac{\varepsilon}{10}\Big) \cap \Omega_3(z) \tag{8.32}$$

if $\sqrt{\kappa+\eta} > N^{-\varepsilon}\hat{\Lambda}$, respectively,

$$\Lambda \leq \frac{N^{\frac{5}{2}\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \qquad \text{on} \quad \Theta\Big(z, \frac{N^{3\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{3\varepsilon}}{\sqrt{N\eta}}\Big) \cap \Omega_3(z) \tag{8.33}$$

if $\sqrt{\kappa+\eta} \leq N^{-\varepsilon}\hat{\Lambda}$. Further, applying (8.32) and (8.33) to (8.25), we can also conclude that

$$\Lambda_{\rm d}(z) \leq \frac{N^{\frac{5}{2}\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \qquad \widetilde{\Lambda}_{\rm d}(z) \leq \frac{N^{\frac{5}{2}\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \qquad \Lambda_T(z) \leq \frac{N^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \qquad \widetilde{\Lambda}_T(z) \leq \frac{N^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}} \tag{8.34}$$

on $\Theta_>\Big(z, \frac{N^{3\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{3\varepsilon}}{\sqrt{N\eta}}, \frac{\varepsilon}{10}\Big) \cap \Omega(z)$ if $\sqrt{\kappa+\eta} > N^{-\varepsilon}\hat{\Lambda}$, and on $\Theta\Big(z, \frac{N^{3\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{3\varepsilon}}{\sqrt{N\eta}}\Big) \cap \Omega(z)$ if $\sqrt{\kappa+\eta} \leq N^{-\varepsilon}\hat{\Lambda}$, where

$$\Omega(z) := \Omega_1(z) \cap \Omega_3(z).$$

Combining (8.32), (8.33) and (8.34), we conclude the proof of Lemma 8.3. □

With Lemma 8.3, we can now prove Theorem 8.1 by using a continuity argument.

Proof of Theorem 8.1. We start with an entry-wise Green function subordination estimate on global scale, *i.e.*, $\eta = \eta_{\mathrm{M}}$ for some sufficiently large constant $\eta_{\mathrm{M}} > 0$. Recall Q_i from (4.11). We regard Q_i as a function of the random unitary matrix U. Then, for $z = E + \mathrm{i}\widetilde{\eta}_M$ with any fixed E and any $\widetilde{\eta}_M \geq \eta_{\mathrm{M}}$, we apply the Gromov-Milman concentration inequality (*cf.*, (6.2) in [6]), and get

$$|Q_i(E + \mathrm{i}\widetilde{\eta}_M) - \mathbb{E}Q_i(E + \mathrm{i}\widetilde{\eta}_M)| \prec \frac{1}{\sqrt{N\widetilde{\eta}_M^4}}\,; \tag{8.35}$$

see Section 6.2 of [6] for similar estimates for the Green function entries of the block additive model.

Next, using the invariance of the Haar measure, one can check the equation

$$\mathbb{E}(\widetilde{B}G \otimes G - G \otimes \widetilde{B}G) = 0\,; \tag{8.36}$$

see Proposition 3.2 of [25]. Taking the (i,i)-th entry for the first component and the normalized trace for the second component in the tensor product, we obtain from (8.36) that

$$\mathbb{E}Q_i = \mathbb{E}\big((\widetilde{B}G)_{ii}\mathrm{tr}\, G - G_{ii}\mathrm{tr}\, \widetilde{B}G\big) = 0\,. \tag{8.37}$$

We claim that, for sufficiently large $\eta_{\mathrm{M}} > 1$, we have

$$\sup_{z:\mathrm{Im}\, z \geq \eta_{\mathrm{M}}} |Q_i(z)| \prec \frac{1}{\sqrt{N}}\,, \qquad \forall i \in [\![1, N]\!]\,, \tag{8.38}$$

where we used (8.35), (8.37), the Lipschitz continuity of Q_i in the regime $|z| \leq \sqrt{N}$ and the deterministic bound $|Q_i(z)| \leq \frac{C}{\sqrt{N}}$ when $|z| \geq \sqrt{N}$. In addition, using that $\|H\| \leq \|A\| + \|B\| < \mathcal{K}$ and the convention $\mathrm{tr}\, \widetilde{B} = \mathrm{tr}\, B = 0$ (*cf.*, (5.1)), we have, for $z = E + \mathrm{i}\widetilde{\eta}_M$ with fixed E and any $\widetilde{\eta}_M \geq \eta_{\mathrm{M}}$, the expansions

$$\mathrm{tr}\, G(z) = -\frac{1}{z} + O(\frac{1}{|z|^2}) = \frac{\mathrm{i}}{\widetilde{\eta}_M} + O\big(\frac{1}{\widetilde{\eta}_M^2}\big), \qquad \mathrm{tr}\, \widetilde{B}G(z) = -\frac{\mathrm{tr}\, \widetilde{B}}{z} + O(\frac{1}{|z|^2}) = O(\frac{1}{\widetilde{\eta}_M^2})\,, \tag{8.39}$$

where we used $\mathrm{tr}\, B = 0$ in the second equality. Hence, by the definition of ω_B^{c} in (5.2), we see that,

$$\omega_B^{\mathrm{c}}(z) = z + O(\frac{1}{\widetilde{\eta}_M}), \qquad z = E + \mathrm{i}\widetilde{\eta}_M. \tag{8.40}$$

Using the identity $(\widetilde{B}G)_{ii} = 1 - (a_i - z)G_{ii}$, we can rewrite (8.38) as

$$(1-(a_i-\omega_B^c)G_{ii})\operatorname{tr} G = O_\prec(\frac{1}{\sqrt{N}}), \qquad z = E + \mathrm{i}\widetilde{\eta}_M .$$

From the first line of (8.39) and (8.40) we get

$$\Lambda_{\mathrm{d}}^c(z) \prec \frac{1}{\sqrt{N}}, \qquad z = E + \mathrm{i}\widetilde{\eta}_M . \tag{8.41}$$

Analogously, we also have

$$\widetilde{\Lambda}_{\mathrm{d}}^c(z) \prec \frac{1}{\sqrt{N}}, \qquad z = E + \mathrm{i}\widetilde{\eta}_M . \tag{8.42}$$

Averaging over the index i in the definition of Λ_{di}^c and $\widetilde{\Lambda}_{di}^c$ (*cf.*, (5.7)), using (8.41) and (8.42) and using the fact $\operatorname{tr} G = \operatorname{tr} \mathcal{G} = m_H$ yields

$$\sup_{z:\operatorname{Im} z \geq \eta_{\mathrm{M}}} \big| m_H(z) - m_A(\omega_B^c(z)) \big| \prec \frac{1}{\sqrt{N}}, \qquad \sup_{z:\operatorname{Im} z \geq \eta_{\mathrm{M}}} \big| m_H(z) - m_B(\omega_A^c(z)) \big| \prec \frac{1}{\sqrt{N}} \tag{8.43}$$

where in the large z regime these bounds even hold deterministically, similarly to (8.38). This together with (5.4) gives us the system

$$\sup_{z:\operatorname{Im} z \geq \eta_{\mathrm{M}}} |\Phi_1(\omega_A^c(z), \omega_B^c(z), z)| \prec \frac{1}{\sqrt{N}}, \qquad \sup_{z:\operatorname{Im} z \geq \eta_{\mathrm{M}}} |\Phi_2(\omega_A^c(z), \omega_B^c(z), z)| \prec \frac{1}{\sqrt{N}}, \tag{8.44}$$

where Φ_1 and Φ_2 are defined in (7.8). We regard (8.44) as a perturbation of $\Phi_1(\omega_A(z), \omega_B(z), z) = 0$, $\Phi_2(\omega_A(z), \omega_B(z), z) = 0$. The stability of this system in the large η regime is analyzed in Lemma A.2. Choosing $(\mu_1, \mu_2) = (\mu_A, \mu_B)$, $(\widetilde{\omega}_1(z), \widetilde{\omega}_2(z)) = (\omega_A^c(z), \omega_B^c(z))$ in Lemma A.2 below, and using the fact that (8.44) and (8.40) hold for any sufficiently large $\widetilde{\eta}_M$, we obtain from the stability Lemma A.2 that

$$|\Lambda_\iota(z)| = |\omega_\iota^c(z) - \omega_\iota(z)| \prec \frac{1}{\sqrt{N}}, \qquad \iota = A, B, \qquad z = E + \mathrm{i}\eta_{\mathrm{M}}, \tag{8.45}$$

for any sufficiently large constant $\eta_{\mathrm{M}} > 1$, say.

Substituting (8.45) into (8.41) and (8.42) gives

$$\Lambda_{\mathrm{d}}(E + \mathrm{i}\eta_{\mathrm{M}}) \prec \frac{1}{\sqrt{N}}, \qquad \widetilde{\Lambda}_{\mathrm{d}}(E + \mathrm{i}\eta_{\mathrm{M}}) \prec \frac{1}{\sqrt{N}}, \tag{8.46}$$

for any fixed $E \in \mathbb{R}$. Using the bound $\|G\| \leq \frac{1}{\eta}$ and the inequality $|\boldsymbol{x}^* G \boldsymbol{y}| \leq \|G\| \|\boldsymbol{x}\| \|\boldsymbol{y}\|$, we also get

$$\Lambda_T(E + \mathrm{i}\eta_{\mathrm{M}}) \leq \frac{1}{\eta_{\mathrm{M}}}, \qquad \widetilde{\Lambda}_T(E + \mathrm{i}\eta_{\mathrm{M}}) \leq \frac{1}{\eta_{\mathrm{M}}}, \tag{8.47}$$

for any fixed $E \in \mathbb{R}$. Since (8.46) and (8.47) guarantee assumption (5.13), we can apply Proposition 5.1 to get, for any fixed $E \in \mathbb{R}$, that

$$\Lambda_T(E + \mathrm{i}\eta_{\mathrm{M}}) \prec \frac{1}{\sqrt{N}}, \qquad \widetilde{\Lambda}_T(E + \mathrm{i}\eta_{\mathrm{M}}) \prec \frac{1}{\sqrt{N}}. \tag{8.48}$$

Also observe that $E + \mathrm{i}\eta_{\mathrm{M}} \in \mathcal{D}_>$, for any fixed E, and that $|\mathcal{S}(E + \mathrm{i}\eta_{\mathrm{M}})| \gtrsim 1$. Hence $\Lambda(E + \mathrm{i}\eta_{\mathrm{M}}) \prec N^{-\varepsilon}|\mathcal{S}(E + \mathrm{i}\eta_{\mathrm{M}})|$. From (8.46), we can also conclude

$$\Lambda(E + \mathrm{i}\eta_{\mathrm{M}}) \prec \frac{1}{\sqrt{N}}. \tag{8.49}$$

Combining (8.46), (8.48), (8.49) with the fact $\Lambda(E + \mathrm{i}\eta_{\mathrm{M}}) \prec N^{-\varepsilon}|\mathcal{S}(E + \mathrm{i}\eta_{\mathrm{M}})|$, we see that the event $\Theta_>(E+\mathrm{i}\eta_{\mathrm{M}}, \frac{N^{3\varepsilon}}{N^{\frac{1}{3}}}, \frac{N^{3\varepsilon}}{\sqrt{N}}, \frac{\varepsilon}{10})$ holds with high probability. More quantitively, we have for any fixed E that

$$\mathbb{P}\Big(\Theta_>(E + \mathrm{i}\eta_{\mathrm{M}}, \frac{N^{3\varepsilon}}{N^{\frac{1}{3}}}, \frac{N^{3\varepsilon}}{\sqrt{N}}, \frac{\varepsilon}{10})\Big) \geq 1 - N^{-D}, \tag{8.50}$$

for all $D > 0$ and $N \geq N_2(D, \varepsilon)$ with some threshold $N_2(D, \varepsilon)$.

Now we take (8.50) as the initial input, and use a continuity argument based on Lemma 8.3, to control the probability of the "good" events $\Theta_>$ for $z \in \mathcal{D}_>$ and Θ for $z \in \mathcal{D}_\leq$. To this end, we first recall the event $\Omega(z)$ in Lemma 8.3. The main task is to show for any $z = E + \mathrm{i}\eta \in \mathcal{D}_>$,

$$\begin{aligned}
&\Theta_>\Big(E + \mathrm{i}\eta, \frac{N^{\frac{5}{2}\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{\frac{5}{2}\varepsilon}}{\sqrt{N\eta}}, \frac{\varepsilon}{2}\Big) \cap \Omega\big(E + \mathrm{i}(\eta - N^{-5})\big) \\
&\quad \subset \Theta_>\Big(E + \mathrm{i}(\eta - N^{-5}), \frac{N^{\frac{5}{2}\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{\frac{5}{2}\varepsilon}}{\sqrt{N\eta}}, \frac{\varepsilon}{2}\Big),
\end{aligned} \tag{8.51}$$

and, for any $z = E + \mathrm{i}\eta \in \mathcal{D}_\leq$,

$$\Theta\Big(E + \mathrm{i}\eta, \frac{N^{\frac{5}{2}\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{\frac{5}{2}\varepsilon}}{\sqrt{N\eta}}\Big) \cap \Omega\big(E + \mathrm{i}(\eta - N^{-5})\big) \subset \Theta\Big(E + \mathrm{i}(\eta - N^{-5}), \frac{N^{\frac{5}{2}\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{\frac{5}{2}\varepsilon}}{\sqrt{N\eta}}\Big). \tag{8.52}$$

The inclusions (8.51) and (8.52) are analogous to (7.20) of [4]. The only difference here is that we decompose the domain $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$ into $\mathcal{D}_>$ and $\mathcal{D}_\leq$, and in $\mathcal{D}_>$ we also keep monitoring the event $\Lambda \leq N^{-\frac{\varepsilon}{2}}|\mathcal{S}|$ in order to use Lemma 8.3 (i). As we are gradually reducing $\operatorname{Im} z$, once z enters into the domain $\mathcal{D}_\leq$, we do not need to monitor $\mathcal{S}$ anymore.

The proofs of (8.51) and (8.52) rely on the Lipschitz continuity of the Green function, $\|G(z) - G(z')\| \leq N^2|z - z'|$, and of the subordination functions and $\mathcal{S}$ in (3.7). Using the Lipschitz continuity of these functions, it is not difficult to see that

$$\Theta_>\Big(E+\mathrm{i}\eta, \frac{N^{\frac{5}{2}\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{\frac{5}{2}\varepsilon}}{\sqrt{N\eta}}, \frac{\varepsilon}{2}\Big) \subset \Theta_>\Big(E+\mathrm{i}(\eta-N^{-5}), \frac{N^{3\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{3\varepsilon}}{\sqrt{N\eta}}, \frac{\varepsilon}{10}\Big),$$
$$z=E+\mathrm{i}\eta\in\mathcal{D}_>\,, \tag{8.53}$$
$$\Theta\Big(E+\mathrm{i}\eta, \frac{N^{\frac{5}{2}\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{\frac{5}{2}\varepsilon}}{\sqrt{N\eta}}\Big) \subset \Theta\Big(E+\mathrm{i}(\eta-N^{-5}), \frac{N^{3\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{3\varepsilon}}{\sqrt{N\eta}}\Big),$$
$$z=E+\mathrm{i}\eta\in\mathcal{D}_\le\,. \tag{8.54}$$

Then, (8.53) together with (8.13) implies (8.51). Similarly, (8.54) together with (8.14) implies (8.52). Applying (8.51) and (8.52) recursively and using the simple fact that the domains $\mathcal{D}_>$ and $\mathcal{D}_\le$ are connected, one can go from $\eta=\eta_{\mathrm{M}}$ to $\eta=\eta_{\mathrm{m}}$, step by step of size N^{-5}. Consequently, we obtain for any $\eta\in[\eta_{\mathrm{m}},\eta_{\mathrm{M}}]\cap N^{-5}\mathbb{Z}$ that, if $E+\mathrm{i}\eta\in\mathcal{D}_>$ then

$$\begin{aligned}&\Theta_>(E+\mathrm{i}\eta_{\mathrm{M}}, \frac{N^{\frac{5}{2}\varepsilon}}{N^{\frac{1}{3}}}, \frac{N^{\frac{5}{2}\varepsilon}}{\sqrt{N}}, \frac{\varepsilon}{2})\cap\Omega(E+\mathrm{i}(\eta_{\mathrm{M}}-N^{-5}))\cap\ldots\cap\Omega(E+\mathrm{i}\eta)\\ &\qquad\subset \Theta_>\Big(E+\mathrm{i}\eta, \frac{N^{\frac{5}{2}\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{\frac{5}{2}\varepsilon}}{\sqrt{N\eta}}, \frac{\varepsilon}{2}\Big)\subset\Theta\Big(E+\mathrm{i}\eta, \frac{N^{\frac{5}{2}\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{\frac{5}{2}\varepsilon}}{\sqrt{N\eta}}\Big),\end{aligned} \tag{8.55}$$

respectively, if $E+\mathrm{i}\eta\in\mathcal{D}_\le$ then

$$\begin{aligned}&\Theta_>(E+\mathrm{i}\eta_{\mathrm{M}}, \frac{N^{\frac{5}{2}\varepsilon}}{N^{\frac{1}{3}}}, \frac{N^{\frac{5}{2}\varepsilon}}{\sqrt{N}}, \frac{\varepsilon}{2})\cap\Omega(E+\mathrm{i}(\eta_{\mathrm{M}}-N^{-5}))\cap\ldots\cap\Omega(E+\mathrm{i}\eta)\\ &\subset\Theta\Big(E+\mathrm{i}\eta, \frac{N^{\frac{5}{2}\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{\frac{5}{2}\varepsilon}}{\sqrt{N\eta}}\Big)\,.\end{aligned} \tag{8.56}$$

Combining (8.12), (8.50), (8.55) and (8.56), we have

$$\mathbb{P}\Big(\Theta\Big(E+\mathrm{i}\eta, \frac{N^{\frac{5}{2}\varepsilon}}{(N\eta)^{\frac{1}{3}}}, \frac{N^{\frac{5}{2}\varepsilon}}{\sqrt{N\eta}}\Big)\Big)\ge 1-N^{-D}(1+N^5(\eta_{\mathrm{M}}-\eta))\,, \tag{8.57}$$

uniformly for all $\eta\in[\eta_{\mathrm{m}},\eta_{\mathrm{M}}]\cap N^{-5}\mathbb{Z}$, when $N\ge\max\{N_1(D,\varepsilon),N_2(D,\varepsilon)\}$. Finally, by the Lipschitz continuity of the Green function and also that of the subordination functions in (3.7), we can extend the bounds from z in the discrete lattice to the entire domain $\mathcal{D}_\tau(\eta_{\mathrm{m}},\eta_{\mathrm{M}})$.

By the definition of the event Θ in (8.9), we obtain from (8.57) that

$$\begin{aligned}&\Lambda_{\mathrm{d}}(z)\le\frac{N^{\frac{5}{2}\varepsilon}}{(N\eta)^{\frac{1}{3}}}\,, \qquad \widetilde{\Lambda}_{\mathrm{d}}(z)\le\frac{N^{\frac{5}{2}\varepsilon}}{(N\eta)^{\frac{1}{3}}}\,, \qquad |\Lambda(z)|\le\frac{N^{\frac{5}{2}\varepsilon}}{(N\eta)^{\frac{1}{3}}}\\ &|\Lambda_T(z)|\le\frac{N^{\frac{5}{2}\varepsilon}}{\sqrt{N\eta}} \qquad |\widetilde{\Lambda}_T(z)|\le\frac{N^{\frac{5}{2}\varepsilon}}{\sqrt{N\eta}}\,,\end{aligned} \tag{8.58}$$

uniformly on $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$ with high probability.

Further, by (8.58), we see that $\varphi(\Gamma_i) = 1$ and $\varphi(\Gamma) = 1$ hold uniformly on $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$, with high probability. Then it is easy to show that the conclusions in (8.23)-(8.25) also hold uniformly on $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$, with high probability. This concludes the proof of Theorem 8.1. $\square$

9. Strong local law

In this section, we will prove the strong local law, *i.e.*, Theorem 2.5. From the weak local law in Theorem 8.1, we have the following rewriting of Proposition 7.1, valid uniformly on $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$.

Proposition 9.1. *Suppose that Assumptions 2.1 and 2.2 hold. Fix any small $\gamma > 0$. Suppose that $\Lambda(z) \prec \hat{\Lambda}(z)$, for some deterministic and positive function $\hat{\Lambda}(z) \prec N^{-\frac{\gamma}{4}}$, then*

$$\Big|\mathcal{S}\Lambda_\iota + \mathcal{T}_\iota\Lambda_\iota^2 + O(\Lambda_\iota^3)\Big| \prec \frac{\sqrt{(\operatorname{Im} m_{\mu_A \boxplus \mu_B} + \hat{\Lambda})(|\mathcal{S}| + \hat{\Lambda})}}{N\eta} + \frac{1}{(N\eta)^2}, \qquad \iota = A, B \tag{9.1}$$

holds uniformly on $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$.

Proof. The proof is the same as Proposition 7.1. But now we have the weak local law Theorem 8.1, which guarantees that the assumptions in Proposition 5.1 hold uniformly on $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$. Hence we do not need additional inputs in (5.13), and the conclusion holds uniformly on $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$. This concludes the proof. $\square$

With the improved bound (9.1) instead of the weaker one in (8.3), we obtain the following improvement of Lemma 8.2.

Lemma 9.2. *Let $\varepsilon \in (0, \frac{\gamma}{100})$. Let $\hat{\Lambda} = \hat{\Lambda}(z) \le N^{-\frac{\gamma}{4}}$ be some deterministic control parameter. Suppose that $\Lambda \le \hat{\Lambda}$. Then we have the following estimates uniformly on $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$:*
(i): If $\sqrt{\kappa + \eta} > N^{-\varepsilon}\hat{\Lambda}$, there is a sufficiently large constant $K_0 > 0$, such that

$$\mathbf{1}\Big(\Lambda \le \frac{|\mathcal{S}|}{K_0}\Big)|\Lambda_A| \prec N^{-2\varepsilon}\hat{\Lambda}\,, \qquad \mathbf{1}\Big(\Lambda \le \frac{|\mathcal{S}|}{K_0}\Big)|\Lambda_B| \prec N^{-2\varepsilon}\hat{\Lambda}\,; \tag{9.2}$$

(ii): If $\sqrt{\kappa + \eta} \le N^{-\varepsilon}\hat{\Lambda}$, we have

$$|\Lambda_A| \prec N^{-\varepsilon}\hat{\Lambda}\,, \qquad |\Lambda_B| \prec N^{-\varepsilon}\hat{\Lambda}. \tag{9.3}$$

Proof. By (9.1) and the fact $\operatorname{Im} m_{\mu_A \boxplus \mu_B} \lesssim |\mathcal{S}|$, we have

$$\Big|\mathcal{S}\Lambda_\iota + \mathcal{T}_\iota\Lambda_\iota^2 + O(\Lambda_\iota^3)\Big| \prec \frac{|\mathcal{S}| + \hat{\Lambda}}{N\eta} + \frac{1}{(N\eta)^2}, \qquad \iota = A, B\,. \tag{9.4}$$

Using (9.4) instead of (8.3), the remaining proof is the same as the proof of Lemma 8.2. $\square$

With the aid of Lemma 9.2, we can now prove (2.17) and (2.18).

Proof of (2.17) and (2.18) in Theorem 2.5. We first prove (2.18). Observe that by the weak local law in Theorem 8.1, we have $\Lambda \prec \frac{1}{(N\eta)^{\frac{1}{3}}}$. Hence, for any $z \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$ we can start with choosing $\hat{\Lambda}(z) = \frac{N^{3\varepsilon}}{(N\eta)^{\frac{1}{3}}} \ll N^{-\frac{\gamma}{4}}$ and apply Lemma 9.2. Now, if $z \in \mathcal{D}_>$ (*cf.* (8.10)), we can use (9.2) iteratively (but for finitely many, $O(\varepsilon^{-1})$ times) to conclude that $\Lambda(z) \prec \frac{1}{N\eta}$. For $z \in \mathcal{D}_\leq$, if $\sqrt{\kappa+\eta} \leq \frac{N^{2\varepsilon}}{N\eta}$, we use (9.3) iteratively until we get $\Lambda(z) \prec \frac{1}{N\eta}$. If $z \in \mathcal{D}_\leq$ and $\sqrt{\kappa+\eta} > \frac{N^{2\varepsilon}}{N\eta}$, we shall first use (9.3) iteratively until we get a bound $\hat{\Lambda}$ which satisfies $\sqrt{\kappa+\eta} > N^{-\varepsilon}\hat{\Lambda}$, then we use (9.2) for the iteration until we get $\Lambda(z) \prec \frac{1}{N\eta}$. Using (7.5) we can get (2.18).

Next, with the weak local law in Theorem 8.1, it is also easy to see that Proposition 6.1 holds uniformly on $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$. For any deterministic $d_1, \ldots, d_N \in \mathbb{C}$, we further write

$$\frac{1}{N}\sum_{i=1}^N d_i\Big(G_{ii} - \frac{1}{a_i - \omega_B^c}\Big) = \frac{1}{N}\sum_{i=1}^N \frac{d_i}{\operatorname{tr} G(a_i - \omega_B^c)} Q_i\,, \tag{9.5}$$

which can easily be checked from the definition of ω_B^c, Q_i and the equation $(a_i - z)G_{ii} + (\widetilde{B}G)_{ii} = 1$. Regarding $\frac{d_i}{\operatorname{tr} G(a_i - \omega_B^c)}$ as the random coefficients d_i in (6.3), it is not difficult to check that (6.2) holds, similarly to the last two equations in (5.56). Hence, we have by Proposition 6.1 that

$$\Big|\frac{1}{N}\sum_{i=1}^N d_i\Big(G_{ii} - \frac{1}{a_i - \omega_B^c}\Big)\Big| \prec \Psi\hat{\Pi}\,. \tag{9.6}$$

Then combining the estimate $\Lambda(z) \prec \frac{1}{N\eta}$ with (9.6) implies (2.17). This concludes the proof of (2.17) and (2.18) in (2.5). $\square$

10. Rigidity of the eigenvalues

In this section, we prove Theorem 2.6, and also (2.19) in Theorem 2.5. We first decompose the domain $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$ into the following two disjoint parts. Fix a small $\epsilon > 0$ and set

$$\begin{aligned}\widehat{\mathcal{D}}_> &:= \Big\{z \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}}) : \sqrt{\kappa+\eta} > \frac{N^{2\varepsilon}}{N\eta}\Big\},\\ \widehat{\mathcal{D}}_\leq &:= \Big\{z \in \mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}}) : \sqrt{\kappa+\eta} \leq \frac{N^{2\varepsilon}}{N\eta}\Big\}.\end{aligned} \tag{10.1}$$

We start by improving the estimate of Λ defined in (7.1) in the following subdomain of $\widehat{\mathcal{D}}_>$,

$$\widetilde{\mathcal{D}}_> := \{z = E + \mathrm{i}\eta \in \widehat{\mathcal{D}}_> : E < E_-\}\,, \tag{10.2}$$

where E_- is the lower endpoint of the support of the measure $\mu_\alpha \boxplus \mu_\beta$; see (2.12).

Lemma 10.1. *Suppose that the assumptions in Theorem 2.5 hold. Then, we have the following uniform estimate for all $z \in \widetilde{\mathcal{D}}_>$,*

$$\Lambda(z) \prec \frac{1}{N\sqrt{(\kappa+\eta)\eta}} + \frac{1}{\sqrt{\kappa+\eta}}\frac{1}{(N\eta)^2}\,. \tag{10.3}$$

Proof. First, from (8.58), we see that $\Lambda \prec \frac{1}{N\eta}$ on $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$. Now, suppose that $\Lambda \prec \hat{\Lambda}$ for some deterministic $\hat{\Lambda} \equiv \hat{\Lambda}(z)$ that satisfies

$$N^\varepsilon\Big(\frac{1}{N\sqrt{(\kappa+\eta)\eta}} + \frac{1}{\sqrt{\kappa+\eta}}\frac{1}{(N\eta)^2}\Big) \le \hat{\Lambda}(z) \le \frac{N^\varepsilon}{N\eta}\,. \tag{10.4}$$

Observe that such $\hat{\Lambda}$ always exists on $\widehat{\mathcal{D}}_>$. From (7.2), (3.4) and (3.5), we have for $\iota = A, B$, and $z \in \widetilde{\mathcal{D}}_>$,

$$\Big|\mathcal{S}\Lambda_\iota + \mathcal{T}_\iota\Lambda_\iota^2\Big| \prec \frac{\sqrt{(\frac{\eta}{\sqrt{\kappa+\eta}} + \hat{\Lambda})(\sqrt{\kappa+\eta} + \hat{\Lambda})}}{N\eta} + \frac{1}{(N\eta)^2} \prec \frac{\sqrt{\hat{\Lambda}\sqrt{\kappa+\eta}}}{N\eta} + \frac{\sqrt{\eta}}{N\eta} + \frac{1}{(N\eta)^2}\,, \tag{10.5}$$

where we used that $\hat{\Lambda} \prec \frac{N^\varepsilon}{N\eta} \le N^{-\varepsilon}\sqrt{\kappa+\eta}$ for all $z \in \widetilde{\mathcal{D}}_>$. Moreover, for $z \in \widetilde{\mathcal{D}}_>$, we see that

$$|\Lambda_\iota| \prec \frac{1}{N\eta} \le N^{-2\varepsilon}\sqrt{\kappa+\eta} \sim N^{-2\varepsilon}|\mathcal{S}|\,,$$

for $\iota = A, B$. Hence, according to the fact $\mathcal{T}_\iota \le C$ (*cf.*, (3.5)), we can absorb the second term on the left side of (10.5) into the first term, and thus we have for $\iota = A, B$

$$|\Lambda_\iota| \prec \frac{1}{\sqrt{\kappa+\eta}}\Bigg(\frac{\sqrt{\hat{\Lambda}\sqrt{\kappa+\eta}}}{N\eta} + \frac{\sqrt{\eta}}{N\eta} + \frac{1}{(N\eta)^2}\Bigg) \le \frac{1}{N\eta(\kappa+\eta)^{\frac14}}\hat{\Lambda}^{\frac12} + N^{-\varepsilon}\hat{\Lambda} \le N^{-\frac{\varepsilon}{4}}\hat{\Lambda}\,,$$

where in the second step we used the lower bound in (10.4) directly, and in the last step we used the fact $(N\eta)^{-1}(\kappa+\eta)^{-\frac14} \le N^{-\frac{\varepsilon}{2}}\hat{\Lambda}^{\frac12}$ which again follows from the lower bound in (10.4).

Hence, we improved the bound from $\Lambda \le \hat{\Lambda}$ to $\Lambda \le N^{-\frac{\varepsilon}{4}}\hat{\Lambda}$ as long as the lower bound in (10.4) holds. Performing the above improvement iteratively, one finally gets (10.3). Hence, we complete the proof. $\square$

With the aid of Lemma 10.1, we can now prove Theorem 2.6.

Proof of Theorem 2.6. We first show (2.21) for the smallest eigenvalue λ_1, *i.e.*,

$$|\lambda_1 - \gamma_1| \prec N^{-\frac{2}{3}} . \tag{10.6}$$

Recall $\mathcal{K}$ defined in (2.13). For any (small) constant $\varepsilon > 0$, we define the line segment.

$$\widetilde{\mathcal{D}}(\varepsilon) := \{z = E + \mathrm{i}\eta : E \in [-\mathcal{K}, E_- - N^{-\frac{2}{3}+6\varepsilon}] , \eta = N^{-\frac{2}{3}+\varepsilon}\}. \tag{10.7}$$

Then it is easy to check that $\widetilde{\mathcal{D}}(\varepsilon) \subset \widetilde{\mathcal{D}}_>$ (*cf.*, (10.2)). Applying (10.3), we obtain $\Lambda \prec \frac{N^{-\varepsilon}}{N\eta}$ uniformly on $\widetilde{\mathcal{D}}(\varepsilon)$, which together with (7.5) implies

$$|m_H(z) - m_{\mu_A \boxplus \mu_B}(z)| \prec \frac{N^{-\varepsilon}}{N\eta} , \tag{10.8}$$

uniformly on $\widetilde{\mathcal{D}}(\varepsilon)$. Moreover, by (3.4), we have

$$\operatorname{Im} m_{\mu_A \boxplus \mu_B}(z) \sim \frac{\eta}{\sqrt{\kappa + \eta}} \le \frac{N^{-\varepsilon}}{N\eta} , \tag{10.9}$$

uniformly on $\widetilde{\mathcal{D}}(\varepsilon)$. Combining (10.8) with (10.9) yields

$$\operatorname{Im} m_H(z) \prec \frac{N^{-\varepsilon}}{N\eta} , \tag{10.10}$$

uniformly on $\widetilde{\mathcal{D}}(\varepsilon)$. Since $\|H\| < \mathcal{K}$, to see (10.6), it suffices to show that with high probability λ_1 is not in the interval $[-\mathcal{K}, E_- - N^{-\frac{2}{3}+6\varepsilon}]$. We prove it by contradiction. Suppose that $\lambda_1 \in [-\mathcal{K}, E_- - N^{-\frac{2}{3}+6\varepsilon}]$. Then clearly for any $\eta > 0$,

$$\sup_{E \in [-\mathcal{K}, E_- - N^{-\frac{2}{3}+6\varepsilon}]} \operatorname{Im} m_H(E + \mathrm{i}\eta) = \sup_{E \in [-\mathcal{K}, E_- - N^{-\frac{2}{3}+6\varepsilon}]} \frac{1}{N} \sum_{i=1}^{N} \frac{\eta}{(\lambda_i - E)^2 + \eta^2} \ge \frac{1}{N\eta} ,$$

which contradicts the fact that (10.10) holds uniformly on $\widetilde{\mathcal{D}}(\varepsilon)$. Hence, we have (10.6).

Next, from (2.18), (3.82) and (3.83) and a standard application of Helffer-Sjöstrand formula (*cf.*, Lemma 5.1 [2]) on $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$ yields

$$\sup_{x \le E_- + c} |\mu_H((-\infty, x]) - \mu_A \boxplus \mu_B((-\infty, x])| \prec \frac{1}{N} , \tag{10.11}$$

for any sufficiently small $c = c(\tau)$. Then (10.6), (10.11), together with the rigidity (3.104) and the square root behavior of the distribution $\mu_\alpha \boxplus \mu_\beta$ (*cf.*, (3.63)) will lead to the conclusion. The same conclusion holds with γ_j^*'s replaced by γ_j's by rigidity (3.104). $\square$

Finally, with the aid of Theorem 2.6, we can prove (2.19) in Theorem 2.5.

Proof of (2.19) in Theorem 2.5. Let $\varepsilon > 0$ be any (small) constant. Since $\kappa = E_- - E \geq N^{-\frac{2}{3}+\varepsilon}$ in (2.19), we see that (2.19) follows from (2.18) directly in the regime $\eta \geq \frac{\kappa}{4}$, say. Hence, in the sequel, we work in the regime $\eta \leq \frac{\kappa}{4}$ only. For any $z = E + \mathrm{i}\eta \in \mathcal{D}_\tau(\eta_\mathrm{m}, \eta_\mathrm{M})$ with $\kappa \geq N^{-\frac{2}{3}+\varepsilon}$, we introduce the contour

$$\mathcal{C} \equiv \mathcal{C}(z) := \mathcal{C}_l \cup \mathcal{C}_r \cup \mathcal{C}_u \cup \overline{\mathcal{C}}_u\,,$$

where

$$\begin{aligned}
\mathcal{C}_l \equiv \mathcal{C}_l(z) &:= \Big\{\tilde{z} = E + \frac{\kappa}{2} + \mathrm{i}\tilde{\eta} : -\eta - \kappa \leq \tilde{\eta} \leq \eta + \kappa\Big\}\,,\\
\mathcal{C}_r \equiv \mathcal{C}_r(z) &:= \Big\{\tilde{z} = E - \frac{\kappa}{2} + \mathrm{i}\tilde{\eta} : -\eta - \kappa \leq \tilde{\eta} \leq \eta + \kappa\Big\}\,,\\
\mathcal{C}_u \equiv \mathcal{C}_u(z) &:= \Big\{\tilde{z} = \tilde{E} + \mathrm{i}(\eta + \kappa) : E - \frac{\kappa}{2} \leq \tilde{E} \leq E + \frac{\kappa}{2}\Big\}\,.
\end{aligned}$$

We then further decompose $\mathcal{C} = \mathcal{C}_< \cup \mathcal{C}_\geq$, where

$$\mathcal{C}_< \equiv \mathcal{C}_<(z) := \big\{\tilde{z} \in \mathcal{C} : |\mathrm{Im}\,\tilde{z}| < \eta_\mathrm{m}\big\}, \qquad \mathcal{C}_\geq \equiv \mathcal{C}_\geq(z) := \mathcal{C} \setminus \mathcal{C}_<.$$

Now, we further introduce the event

$$\Xi := \bigcap_{\tilde{z}\in\mathcal{C}>} \Big\{\big|m_H(\tilde{z}) - m_{\mu_A\boxplus\mu_B}(\tilde{z})\big| \leq \frac{N^\varepsilon}{N\mathrm{Im}\,\tilde{z}}\Big\} \bigcap \Big\{\lambda_1 \geq E_- - \frac{1}{4}N^{-2/3+\varepsilon}\Big\}\,.$$

Then, on the event Ξ, we have

$$\begin{aligned}
m_H(z) - m_{\mu_A\boxplus\mu_B}(z) &= \frac{1}{2\pi\mathrm{i}}\oint_{\mathcal{C}} \frac{1}{\tilde{z} - z}\big(m_H(\tilde{z}) - m_{\mu_A\boxplus\mu_B}(\tilde{z})\big)\mathrm{d}\tilde{z}\\
&= \frac{1}{2\pi\mathrm{i}}\Big(\int_{\mathcal{C}_<} + \int_{\mathcal{C}_\geq}\Big)\frac{1}{\tilde{z} - z}\big(m_H(\tilde{z}) - m_{\mu_A\boxplus\mu_B}(\tilde{z})\big)\mathrm{d}\tilde{z}. \qquad (10.12)
\end{aligned}$$

Note that, for $\tilde{z} \in \mathcal{C}$, we always have $\frac{1}{|\tilde{z}-z|} \leq \frac{2}{\kappa}$. In addition, for $\tilde{z} \in \mathcal{C}_<$, we have the fact $|\mathcal{C}_<| \leq \eta_\mathrm{m}$, and

$$|m_H(\tilde{z})| \leq \frac{C}{\kappa}\,, \qquad |m_{\mu_A\boxplus\mu_B}(\tilde{z})| \leq \frac{C}{\kappa}\,,$$

which hold on Ξ. For $\tilde{z} \in \mathcal{C}_\geq$, we have the fact $|\mathcal{C}_\geq| \leq C\kappa$ and the bound

$$\big|m_H(\tilde{z}) - m_{\mu_A\boxplus\mu_B}(\tilde{z})\big| \leq \frac{N^\varepsilon}{N\mathrm{Im}\,\tilde{z}}\,,$$

which holds on Ξ. Applying the above bounds to (10.12), it is elementary to check that

$$|m_H(z) - m_{\mu_A \boxplus \mu_B}(z)| \le C\big(\eta_{\mathrm{m}} + N^{-1+\varepsilon}\log N\big)\frac{1}{\kappa}$$

on Ξ. Since γ in $\eta_{\mathrm{m}} = N^{-1+\gamma}$ and ε can be arbitrary, we can conclude that

$$|m_H(z) - m_{\mu_A \boxplus \mu_B}(z)| \prec \frac{1}{N\kappa} \tag{10.13}$$

if we can show that Ξ holds with high probability. Using (10.6), it suffices to show that

$$\big|m_H(\tilde{z}) - m_{\mu_A \boxplus \mu_B}(\tilde{z})\big| \prec \frac{1}{N\mathrm{Im}\,\tilde{z}}\,,$$

uniformly in $\tilde{z} \in \mathcal{C}_>$. This only requires enlarging the domain $\mathcal{D}_\tau(\eta_{\mathrm{m}}, \eta_{\mathrm{M}})$ and also consider its complex conjugate to include $\mathcal{C}_>$ during the proof of (2.18). Hence, we conclude the proof of (2.19) by combining the $\frac{1}{N\kappa}$ bound in (10.13) with the $\frac{1}{N\eta}$ bound in (2.18). □

We conclude the main part of the paper with the proof of Corollary 2.8.

Proof of Corollary 2.8. With the additional Assumption 2.7, we can show analogously that the estimates (2.18) and (2.21) hold as well around the upper edge. According to Assumption 2.7 (*vii*) and the fact $\sup_{\mathbb{C}^+} |m_{\mu_\alpha \boxplus \mu_\beta}| \le C$ (*cf.*, (3.8)), we see that except for the two vicinities of the lower and upper edge, the remaining spectrum is within the regular bulk. Together with the strong local law in the bulk regime, *cf.*, Theorem 2.4 in [5], we have

$$\big|m_H(z) - m_{\mu_A \boxplus \mu_B}(z)\big| \prec \frac{1}{N\eta}, \tag{10.14}$$

uniformly on the domain $\mathcal{D}(\eta_{\mathrm{m}}, \eta_{\mathrm{M}}) := \{z = E + \mathrm{i}\eta \in \mathbb{C}^+ : -\mathcal{K} \le E \le \mathcal{K}, \quad \eta_{\mathrm{m}} \le \eta \le \eta_{\mathrm{M}}\}$. Then, (10.14) together with (2.21) and its counterpart at the upper edge implies the rigidity for all eigenvalues, *i.e.*, (2.22) can be proved again with Helffer-Sjöstrand formula. Then, from (2.22), we conclude that (2.23) holds. This completes the proof of Corollary 2.8. □

Acknowledgments

The authors are grateful to the anonymous referee for pointing out a gap in the first version of the paper as well as for numerous useful remarks and comments.

Appendix A

In this appendix, we collect some basic technical results.

A.1. Stochastic domination and large deviation properties

Recall the stochastic domination in Definition 2.4. The relation $\prec$ is transitive and it satisfies the following arithmetic rules: if $X_1 \prec Y_1$ and $X_2 \prec Y_2$ then $X_1 + X_2 \prec Y_1 + Y_2$ and $X_1 X_2 \prec Y_1 Y_2$. Further assume that $\Phi(v) \geq N^{-C}$ is deterministic and that $Y(v)$ is a nonnegative random variable satisfying $\mathbb{E}[Y(v)]^2 \leq N^{C'}$ for all v. Then $Y(v) \prec \Phi(v)$, uniformly in v, implies $\mathbb{E}[Y(v)] \prec \Phi(v)$, uniformly in v.

Gaussian vectors have well-known large deviation properties which we use in the following form:

Lemma A.1. *Let $X = (x_{ij}) \in M_N(\mathbb{C})$ be a deterministic matrix and let $\boldsymbol{y} = (y_i) \in \mathbb{C}^N$ be a deterministic complex vector. For a Gaussian random vector $\boldsymbol{g} = (g_1, \ldots, g_N) \in \mathcal{N}_{\mathbb{R}}(0, \sigma^2 I_N)$ or $\mathcal{N}_{\mathbb{C}}(0, \sigma^2 I_N)$, we have*

$$|\boldsymbol{y}^* \boldsymbol{g}| \prec \sigma \|\boldsymbol{y}\|\,, \qquad\qquad |\boldsymbol{g}^* X \boldsymbol{g} - \sigma^2 N \operatorname{tr} X| \prec \sigma^2 \|X\|_2\,. \tag{A.1}$$

A.2. Stability for large η

For any probability measures μ_1 and μ_2 on the real line, we define the functions $\Phi_1,\ \Phi_2 : (\mathbb{C}^+)^3 \to \mathbb{C}$ by setting

$$\Phi_1(\omega_1, \omega_2, z) := F_{\mu_1}(\omega_2) - \omega_1 - \omega_2 + z\,, \qquad \Phi_2(\omega_1, \omega_2, z) := F_{\mu_2}(\omega_1) - \omega_1 - \omega_2 + z\,. \tag{A.2}$$

We observe that the system of subordination equations (2.9) is equivalent to

$$\Phi_1(\omega_1(z), \omega_2(z), z) = 0\,, \qquad \Phi_1(\omega_1(z), \omega_2(z), z) = 0\,, \qquad\qquad \forall z \in \mathbb{C}^+.$$

We have the following linear stability for the subordination equation in the large η regime. A somewhat weaker version of this result has already been proven in Lemma 4.2 of [3] requiring an unnecessarily stronger condition (compare (4.14) of [3] with the current (A.3) below). However, in our applications only a weaker assumption can be guaranteed. In fact, already in [3] (in equation (6.56)) we tacitly relied on the current version of this stability result. Thus by proving the stronger stability result below we also correct this small inconsistency in [3].

Lemma A.2. *Let $\widetilde{\eta}_0 > 0$ be any (large) positive number and let $\widetilde{\omega}_1, \widetilde{\omega}_2, \widetilde{r}_1, \widetilde{r}_2 : \mathbb{C}_{\widetilde{\eta}_0} \to \mathbb{C}$ be analytic functions where $\mathbb{C}_{\widetilde{\eta}_0} := \{z \in \mathbb{C} \;:\; \operatorname{Im} z \geq \widetilde{\eta}_0\}$. Assume that there is a constant $C > 0$ such that the following hold for all $z \in \mathbb{C}_{\widetilde{\eta}_0}$:*

$$|\operatorname{Im} \widetilde{\omega}_1(z) - \operatorname{Im} z| \leq C\,, \qquad\qquad |\operatorname{Im} \widetilde{\omega}_2(z) - \operatorname{Im} z| \leq C\,, \tag{A.3}$$

$$|\widetilde{r}_1(z)| \leq C\,, \qquad\qquad |\widetilde{r}_2(z)| \leq C\,, \tag{A.4}$$

$$\Phi_1(\widetilde{\omega}_1(z), \widetilde{\omega}_2(z), z) = \widetilde{r}_1(z)\,, \qquad\qquad \Phi_2(\widetilde{\omega}_1(z), \widetilde{\omega}_2(z), z) = \widetilde{r}_2(z)\,. \tag{A.5}$$

Then there is a constant η_0 with $\eta_0 \geq \widetilde{\eta}_0$, such that

$$|\widetilde{\omega}_1(z) - \omega_1(z)| \leq 2\|\widetilde{r}(z)\|\,, \qquad\qquad |\widetilde{\omega}_2(z) - \omega_2(z)| \leq 2\|\widetilde{r}(z)\|\,, \tag{A.6}$$

on the domain $\mathbb{C}_{\eta_0} := \{z \in \mathbb{C} \,:\, \operatorname{Im} z \geq \eta_0\}$, where $\omega_1(z)$ and $\omega_2(z)$ are the subordination functions associated with μ_1 and μ_2.

Proof. Since most of the proof is identical to that in [3], here we only give the necessary modifications involving the weaker condition (A.3). Following the proof in [3] to the letter up to (4.23), for every $z \in \mathbb{C}_{\eta_0}$ we have constructed functions $\widehat{\omega}_1(z)$, $\widehat{\omega}_2(z)$ such that $\Phi_{\mu_1,\mu_2}(\widehat{\omega}_1(z), \widehat{\omega}_2(z), z) = 0$ with

$$|\widetilde{\omega}_j(z) - \widehat{\omega}_j(z)| \leq 2\|\widetilde{r}(z)\|\,, \qquad\qquad j = 1, 2\,, \qquad z \in \mathbb{C}_{\eta_0}\,. \tag{A.7}$$

From (4.20) of [3] we know that the Jacobian of the subordination equations (denoted by Γ_{μ_1,μ_2} in [3]) is close to 1 for sufficiently large $\widetilde{\eta}_0$. Thus by analytic inverse function theorem we obtain that $\widehat{\omega}_j(z)$, $j = 1, 2$, are also analytic functions for large $\eta = \operatorname{Im} z$. From (A.3), (A.4) and (A.7), we see that

$$\lim_{\eta \nearrow \infty} \frac{\operatorname{Im} \widehat{\omega}_1(\mathrm{i}\eta)}{\mathrm{i}\eta} = \lim_{\eta \nearrow \infty} \frac{\operatorname{Im} \widehat{\omega}_2(\mathrm{i}\eta)}{\mathrm{i}\eta} = 1\,.$$

It is known from the proof of the uniqueness of the solution to the subordination equations near $z = \mathrm{i}\infty$ that $(\widehat{\omega}_1(z), \widehat{\omega}_2(z))$ is the unique solution in a neighborhood of $z = \mathrm{i}\infty$ and it can be analytically extended to all $z \in \mathbb{C}^+$. Hence, $(\widehat{\omega}_1(z), \widehat{\omega}_2(z)) = (\omega_1(z), \omega_2(z))$. This together with (A.7) concludes the proof. $\square$

Appendix B

In this appendix, we prove some technical lemmas. First, we estimate the small terms involving Δ_G. Specifically, we provide the bounds for the Δ_G involved terms in the last four estimates in Lemma 5.3. Then, we prove Lemma 5.3. We summarize the estimates for Δ_G involved terms in the following lemma.

Lemma B.1. *Fix a $z \in \mathcal{D}_\tau(\eta_\mathrm{m}, \eta_\mathrm{M})$. Let $Q \in M_N(\mathbb{C})$ be arbitrary, with $\|Q\| \prec 1$. Let $X_i = I$ or $\widetilde{B}^{\langle i \rangle}$, and $X = I$ or A. Suppose the assumptions of Proposition 5.1 hold. Then, we have*

$$\frac{1}{N} \sum_k^{(i)} \boldsymbol{e}_k^* X_i \Delta_G(i,k) \boldsymbol{e}_i = O_\prec(\Pi_i^2),$$

$$\frac{1}{N} \sum_k^{(i)} \boldsymbol{e}_i^* X \Delta_G(i,k) \boldsymbol{e}_i \boldsymbol{e}_k^* X_i G \boldsymbol{e}_i = O_\prec(\Pi_i^2),$$

$$\frac{1}{N}\sum_{k}^{(i)} \boldsymbol{h}_i^* \Delta_G(i,k)\boldsymbol{e}_i\boldsymbol{e}_k^* X_i G\boldsymbol{e}_i = O_\prec(\Pi_i^2),$$
$$\frac{1}{N}\sum_{k}^{(i)} \operatorname{tr} QX\Delta_G(i,k)\boldsymbol{e}_k^* X_i G\boldsymbol{e}_i = O_\prec(\Psi^2\Pi_i^2). \tag{B.1}$$

Proof. The proof is similar to that of Lemma B.1 in [6]. But here we need finer estimates. Recall $\Delta_R(i,k)$ and $\Delta_G(i,k)$ from (5.40) and (5.39). We note that $\Delta_R(i,k)$ is a sum of terms of the form $\widetilde{d}_i \bar{g}_{ik}\boldsymbol{\alpha}_i\boldsymbol{\beta}_i^*$ for some $\widetilde{d}_i \in \mathbb{C}$ with $|\widetilde{d}_i| \prec 1$, where $\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i = \boldsymbol{e}_i$ or $\boldsymbol{h}_i$. Hereafter, we use $\widetilde{d}_i$ to represent a generic number satisfying $|\widetilde{d}_i| \prec 1$ uniformly on $\mathcal{D}_\tau(\eta_{\mathrm{m}}, 1)$. Then, we see that $\Delta_G(i,k)$ is a sum of terms of the form

$$\widetilde{d}_i \bar{g}_{ik} G\boldsymbol{\alpha}_i\boldsymbol{\beta}_i^* \widetilde{B}^{\langle i\rangle} R_i G, \qquad \widetilde{d}_i \bar{g}_{ik} GR_i \widetilde{B}^{\langle i\rangle}\boldsymbol{\alpha}_i\boldsymbol{\beta}_i^* G. \tag{B.2}$$

Then, the left hand side of the first estimate in (B.1) is a sum of terms of the form

$$\frac{1}{N}\widetilde{d}_i\big(\mathring{\boldsymbol{g}}_i^* X_i G\boldsymbol{\alpha}_i\big)\big(\boldsymbol{\beta}_i^*\widetilde{B}^{\langle i\rangle} R_i G\boldsymbol{e}_i\big), \qquad \frac{1}{N}\widetilde{d}_i\big(\mathring{\boldsymbol{g}}_i^* X_i GR_i\widetilde{B}^{\langle i\rangle}\boldsymbol{\alpha}_i\big)\big(\boldsymbol{\beta}_i^* G\boldsymbol{e}_i\big). \tag{B.3}$$

By the Cauchy-Schwarz inequality, we have

$$\big|\mathring{\boldsymbol{g}}_i^* X_i G\boldsymbol{\alpha}_i\big| \prec \|G\boldsymbol{\alpha}_i\| = \sqrt{\frac{\operatorname{Im}\boldsymbol{\alpha}_i^* G\boldsymbol{\alpha}_i}{\eta}},$$
$$\big|\boldsymbol{\beta}_i^*\widetilde{B}^{\langle i\rangle} R_i G\boldsymbol{e}_i\big| \prec \|G\boldsymbol{e}_i\| = \sqrt{\frac{\operatorname{Im} G_{ii}}{\eta}}, \qquad \big|\boldsymbol{\beta}_i^* G\boldsymbol{e}_i\big| \prec \|G\boldsymbol{e}_i\| = \sqrt{\frac{\operatorname{Im} G_{ii}}{\eta}},$$
$$\big|\mathring{\boldsymbol{g}}_i^* X_i GR_i\widetilde{B}^{\langle i\rangle}\boldsymbol{\alpha}_i\big| \prec \|GR_i\widetilde{B}^{\langle i\rangle}\boldsymbol{\alpha}_i\| = \sqrt{\frac{\operatorname{Im}\boldsymbol{\alpha}_i^*\widetilde{B}^{\langle i\rangle} R_i GR_i\widetilde{B}^{\langle i\rangle}\boldsymbol{\alpha}_i}{\eta}}. \tag{B.4}$$

Note that for $\boldsymbol{\alpha}_i = \boldsymbol{e}_i$,

$$\boldsymbol{\alpha}_i^* G\boldsymbol{\alpha}_i = G_{ii}, \qquad \boldsymbol{\alpha}_i^*\widetilde{B}^{\langle i\rangle} R_i GR_i\widetilde{B}^{\langle i\rangle}\boldsymbol{\alpha}_i = b_i^2\boldsymbol{h}_i^* G\boldsymbol{h}_i = b_i^2\mathcal{G}_{ii}, \tag{B.5}$$

and for $\boldsymbol{\alpha}_i = \boldsymbol{h}_i$,

$$\boldsymbol{\alpha}_i^* G\boldsymbol{\alpha}_i = \mathcal{G}_{ii}, \qquad \boldsymbol{\alpha}_i^*\widetilde{B}^{\langle i\rangle} R_i GR_i\widetilde{B}^{\langle i\rangle}\boldsymbol{\alpha}_i = \boldsymbol{e}_i^*\widetilde{B}G\widetilde{B}\boldsymbol{e}_i = \widetilde{B}_{ii} - (a_i - z) + (a_i - z)G_{ii}. \tag{B.6}$$

Plugging (B.5) and (B.6) into the bounds in (B.4), we see that both terms in (B.3) are of order $O_\prec(\Pi_i^2)$. Hence, we proved the first estimate in (B.1).

Next, we verify the second estimate (B.1). Since $\Delta_G(i,k)$ is a sum of terms of the form in (B.2), we see that the left side of the second estimate in (B.1) is a sum of terms of the form

$$\frac{1}{N}\widetilde{d}_i\big(\boldsymbol{e}_i^* XG\boldsymbol{\alpha}_i\big)\big(\boldsymbol{\beta}_i^*\widetilde{B}^{\langle i\rangle}R_iG\boldsymbol{e}_i\big)\big(\mathring{\boldsymbol{g}}_i^* X_iG\boldsymbol{e}_i\big)\,,$$
$$\frac{1}{N}\widetilde{d}_i\big(\boldsymbol{e}_i^* XGR_i\widetilde{B}^{\langle i\rangle}\boldsymbol{\alpha}_i\big)\big(\boldsymbol{\beta}_i^* G\boldsymbol{e}_i\big)\big(\mathring{\boldsymbol{g}}_i^* X_iG\boldsymbol{e}_i\big)\,. \tag{B.7}$$

Note that

$$\boldsymbol{e}_i^*\widetilde{B}^{\langle i\rangle}R_iG\boldsymbol{e}_i=-b_iT_i,\qquad \boldsymbol{h}_i^*\widetilde{B}^{\langle i\rangle}R_iG\boldsymbol{e}_i=-(\widetilde{B}G)_{ii}.$$

Hence, we have

$$|\boldsymbol{\beta}_i^*\widetilde{B}^{\langle i\rangle}R_iG\boldsymbol{e}_i|\prec 1,\qquad |\boldsymbol{\beta}_i^* G\boldsymbol{e}_i|\prec 1. \tag{B.8}$$

Further, we claim that

$$|\boldsymbol{e}_i^* XG\boldsymbol{\alpha}_i|,\ |\boldsymbol{e}_i^* XGR_i\widetilde{B}^{\langle i\rangle}\boldsymbol{\alpha}_i|\prec\sqrt{\frac{\operatorname{Im}(G_{ii}+\mathcal{G}_{ii})}{\eta}}. \tag{B.9}$$

The proof of the above bounds is analogous to the proof of (B.4). We thus omit the details. Then, using the first estimate in (B.4), (B.8) and (B.9), we see that both terms in (B.7) are of order $O_\prec(\Pi_i^2)$.

The proof of the third estimate in (B.1) is nearly the same as that for the second one, we thus omit it.

To show the last estimate, we again use the fact that $\Delta_G(i,k)$ is a sum of terms of the form in (B.2). Then it is not difficult to see that the left side of the last estimate in (B.1) is a sum of terms of the form

$$\frac{\widetilde{d}_i}{N^2}\big(\boldsymbol{\beta}_i^*\widetilde{B}^{\langle i\rangle}R_iGQXG\boldsymbol{\alpha}_i\big)\big(\mathring{\boldsymbol{g}}_i^* X_iG\boldsymbol{e}_i\big),\qquad \frac{\widetilde{d}_i}{N^2}\big(\boldsymbol{\beta}_i^* GQXGR_i\widetilde{B}^{\langle i\rangle}\boldsymbol{\alpha}_i\big)\big(\mathring{\boldsymbol{g}}_i^* X_iG\boldsymbol{e}_i\big). \tag{B.10}$$

Note that

$$\big|\boldsymbol{\beta}_i^*\widetilde{B}^{\langle i\rangle}R_iGQXG\boldsymbol{\alpha}_i\big|\prec\frac{1}{\eta}\|G\boldsymbol{\alpha}_i\|\le\frac{1}{\eta}\sqrt{\frac{\operatorname{Im}(G_{ii}+\mathcal{G}_{ii})}{\eta}}. \tag{B.11}$$

Analogously, we have

$$\big|\boldsymbol{\beta}_i^* GQXGR_i\widetilde{B}^{\langle i\rangle}\boldsymbol{\alpha}_i\big|\prec\frac{1}{\eta}\sqrt{\frac{\operatorname{Im}(G_{ii}+\mathcal{G}_{ii})}{\eta}}. \tag{B.12}$$

Applying (B.11), (B.12), and the first estimate in (B.4), we see that both terms in (B.10) are of order $O_\prec(\Psi^2\Pi_i^2)$. Hence, we obtain the last estimate in (B.1). This concludes the proof of Lemma B.1. $\square$

Proof of Lemma 5.3. The proof is similar to that for Lemma 7.4 in [6]. In the latter, we used Ψ instead of Π_i in the statement. However, the proof of Lemma 7.4 in [6] shows readily that the stronger bounds in (5.56) hold for the counterparts of the block additive model (*cf.*, (7.77), (7.80), (7.81) and (7.87) of [6]). The proof for our additive model given here analogous.

First, by (5.17), (5.18), (5.28), (5.31), and the fact $\mathring{T}_i = T_i - h_{ii}G_{ii}$, we have $|\mathring{S}_i| \prec 1$, $|\mathring{T}_i| \prec 1$, under the assumption ((5.13). Then, for the first estimate in (5.56), we have

$$\frac{1}{N}\sum_k^{(i)} \frac{\partial \|\boldsymbol{g}_i\|^{-1}}{\partial g_{ik}} \boldsymbol{e}_k^* X_i G \boldsymbol{e}_i = -\frac{1}{2N}\frac{1}{\|\boldsymbol{g}_i\|^3}\sum_k^{(i)} \bar{g}_{ik}\boldsymbol{e}_k^* X_i \boldsymbol{e}_i = -\frac{1}{2N}\frac{1}{\|\boldsymbol{g}_i\|^2}\mathring{\boldsymbol{h}}_i^* X_i G\boldsymbol{e}_i$$
$$= O_\prec(\frac{1}{N}),$$

where we used the fact that $\mathring{\boldsymbol{h}}_i^* X_i G\boldsymbol{e}_i = \mathring{S}_i$ or $\mathring{T}_i$ if $X_i = \widetilde{B}^{\langle i\rangle}$ or I, respectively.

Next, we show the second bound in (5.56). It is convenient to set $I^{\langle i\rangle} := I - \boldsymbol{e}_i\boldsymbol{e}_i^*$. Using (5.38), we get

$$\frac{1}{N}\sum_k^{(i)} \boldsymbol{e}_i^* X \frac{\partial G}{\partial g_{ik}}\boldsymbol{e}_i\boldsymbol{e}_k^* X_i G\boldsymbol{e}_i = \frac{c_i}{N}\boldsymbol{e}_i^* XGI^{\langle i\rangle}X_iG\boldsymbol{e}_i(\boldsymbol{e}_i + \boldsymbol{h}_i^*)\widetilde{B}^{\langle i\rangle}R_iG\boldsymbol{e}_i$$
$$+\frac{c_i}{N}\boldsymbol{e}_i^* XGR_i\widetilde{B}^{\langle i\rangle}I^{\langle i\rangle}X_iG\boldsymbol{e}_i(\boldsymbol{e}_i+\boldsymbol{h}_i)^*G\boldsymbol{e}_i + \frac{1}{N}\sum_k^{(i)}\boldsymbol{e}_i^* X\Delta_G(i,k)\boldsymbol{e}_i\boldsymbol{e}_k^* X_iG\boldsymbol{e}_i. \tag{B.13}$$

The desired estimate of the last term was obtained in the second line of (B.1). Further, using (4.8) we get

$$(\boldsymbol{e}_i + \boldsymbol{h}_i^*)\widetilde{B}^{\langle i\rangle}R_iG\boldsymbol{e}_i = -b_iT_i - (\widetilde{B}G)_{ii} = O_\prec(1), \qquad (\boldsymbol{e}_i+\boldsymbol{h}_i)^*G\boldsymbol{e}_i = G_{ii} + T_i = O_\prec(1),$$

where the estimates follow from (5.17) and (5.18). Hence, it suffices to show that

$$|\boldsymbol{e}_i^* XGI^{\langle i\rangle}X_iG\boldsymbol{e}_i| \prec \frac{\mathrm{Im}\,(G_{ii}+\mathcal{G}_{ii})}{\eta}, \qquad |\boldsymbol{e}_i^* XGR_i\widetilde{B}^{\langle i\rangle}I^{\langle i\rangle}X_iG\boldsymbol{e}_i| \prec \frac{\mathrm{Im}\,(G_{ii}+\mathcal{G}_{ii})}{\eta}. \tag{B.14}$$

Note that, by the assumption $X = I$ or A, both terms in (B.14) can be bounded by

$$C\|GX\boldsymbol{e}_i\|\|G\boldsymbol{e}_i\| = \frac{C}{\eta}\sqrt{\mathrm{Im}\,(XGX)_{ii}}\sqrt{\mathrm{Im}\,G_{ii}} \le C'\frac{\mathrm{Im}\,G_{ii}}{\eta}.$$

This completes the proof of the second inequality in (5.56). Next, we show the third estimate in (5.56). In light of the definition of T_i, it suffices to show

$$\frac{1}{N}\sum_{k}^{(i)}\frac{\partial \boldsymbol{h}_i^*}{\partial g_{ik}}G\boldsymbol{e}_i\boldsymbol{e}_k^*X_iG\boldsymbol{e}_i=O_\prec(\frac{1}{N}),\qquad \frac{1}{N}\sum_{k}^{(i)}\boldsymbol{h}_i^*\frac{\partial G}{\partial g_{ik}}\boldsymbol{e}_i\boldsymbol{e}_k^*X_iG\boldsymbol{e}_i=O_\prec(\Pi_i^2).\tag{B.15}$$

The first estimate in (B.15) is proved as follows

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

where in the last step we again use the fact $\mathring{\boldsymbol{h}}_i^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i=\mathring{S}_i=O_\prec(1)$ and $\boldsymbol{h}_i^*G\boldsymbol{e}_i=T_i=O_\prec(1)$. The proof of the second estimate in (B.15) is similar to that of the second inequality in (5.56). It suffices to replace $\boldsymbol{e}_i^*X$ by $\boldsymbol{h}_i^*$ in (B.13) and estimate the resulting terms. The counterpart to the last term in (B.13) is estimated in (B.1). The counterparts to the first two terms on the right side of (B.13) are bounded by

$$C\|G\boldsymbol{h}_i\|\|G\boldsymbol{e}_i\|=\frac{C}{\eta}\sqrt{\operatorname{Im}\boldsymbol{h}_i^*G\boldsymbol{h}_i}\sqrt{\operatorname{Im}G_{ii}}=\frac{C}{\eta}\sqrt{\operatorname{Im}\mathcal{G}_{ii}}\sqrt{\operatorname{Im}G_{ii}}\le C'\frac{\operatorname{Im}(G_{ii}+\mathcal{G}_{ii})}{\eta},$$

where we have used (5.44).

Next, we show the fourth estimate in (5.56). Using (5.38) again, we can get

$$\begin{aligned}\frac{1}{N}\sum_{k}^{(i)}\operatorname{tr}\Big(QX\frac{\partial G}{\partial g_{ik}}\Big)\boldsymbol{e}_k^*X_iG\boldsymbol{e}_i&=\frac{c_i}{N^2}(\boldsymbol{e}_i+\boldsymbol{h}_i)^*\widetilde{B}^{\langle i\rangle}R_iGQXGI^{\langle i\rangle}X_iG\boldsymbol{e}_i\\&\quad+\frac{c_i}{N^2}(\boldsymbol{e}_i+\boldsymbol{h}_i)^*GQXGR_i\widetilde{B}^{\langle i\rangle}I^{\langle i\rangle}X_iG\boldsymbol{e}_i+\frac{1}{N}\sum_{k}^{(i)}\operatorname{tr}QX\Delta_G(i,k)\boldsymbol{e}_k^*X_iG\boldsymbol{e}_i.\end{aligned}\tag{B.16}$$

The last term above is estimated in (B.1). Using (4.8) and $\|G\|\le\eta$, we have

$$\begin{aligned}\Big|\frac{1}{N^2}(\boldsymbol{e}_i+\boldsymbol{h}_i)^*\widetilde{B}^{\langle i\rangle}R_iGQXGI^{\langle i\rangle}X_iG\boldsymbol{e}_i\Big|&=\Big|\frac{1}{N^2}(b_i\boldsymbol{h}_i^*+\boldsymbol{e}_i^*\widetilde{B})GQXGI^{\langle i\rangle}X_iG\boldsymbol{e}_i\Big|\\&\le C\frac{1}{N^2\eta}\big(\|G\boldsymbol{h}_i\|+\|G\widetilde{B}\boldsymbol{e}_i\|\big)\|G\boldsymbol{e}_i\|\le C\frac{1}{N^2\eta}\big(\|G\boldsymbol{h}_i\|^2+\|G\widetilde{B}\boldsymbol{e}_i\|^2+\|G\boldsymbol{e}_i\|^2\big)\\&=\frac{C}{N^2\eta^2}\big(\operatorname{Im}(\boldsymbol{h}_i^*G\boldsymbol{h}_i+(\widetilde{B}G\widetilde{B})_{ii}+G_{ii})\big)\prec\frac{\operatorname{Im}(G_{ii}+\mathcal{G}_{ii})}{N^2\eta^2}.\end{aligned}\tag{B.17}$$

Here in the last step we again used (5.44) and also fact

$$\operatorname{Im}(\widetilde{B}G\widetilde{B})_{ii}=\eta+\operatorname{Im}((a_i-z)^2G_{ii})=O_\prec(\eta+\operatorname{Im}G_{ii})=O_\prec(\operatorname{Im}G_{ii}).\tag{B.18}$$

In (B.18), we used (5.9), the first bound in (5.17), and $\operatorname{Im} G_{ii} \gtrsim \eta$ which is easily checked by spectral decomposition. Similar to (B.17), we get the desired estimate for the second term on the right of (B.16).

Finally, the last equation in (5.56) can be proved analogously to the fourth one. The only difference is, instead of the factor $\boldsymbol{e}_k^* X_i G \boldsymbol{e}_i$ in (6.22), here we have $\boldsymbol{e}_k^* X_i \mathring{\boldsymbol{g}}_i$ which does not contain any G factor, which actually makes the estimates even simpler. This completes the proof of Lemma 5.3. $\square$

Appendix C. Estimates of the cutoff errors

In this appendix, we state more details on the estimate (8.29). The proof can be done in the same way as the non-cutoff version (6.5), but with the a priori inputs given by $\varphi(\Gamma_i)$'s and $\varphi(\Gamma)$. Since the proof can be done via going through the proof of (6.5) again, we only list the necessary modifications here.

The first modification we need to do is the bound of the analogue of the term $O_\prec(\Psi\hat{\Upsilon})$ in (6.6). This error term was obtained when we bounded the term $\frac{1}{N}\sum_{i=1}^N T_i \tau_{i1} \Upsilon$ in (6.19). Here, during the proof of (8.29), the counterpart will be $\frac{1}{N}\sum_{i=1}^N T_i \widetilde{\tau}_{i1} \Upsilon$, where $\widetilde{\tau}_{i1}$ is defined via replacing all d_j's by $d_j \varphi(\Gamma_j)\varphi(\Gamma)$ in the definition of τ_{i1} in (6.12). According to the definition of Γ in (8.27), it is easy to see that

$$\Big|\frac{1}{N}\sum_{i=1}^N T_i \widetilde{\tau}_{i1}\Upsilon\Big| \le C\frac{N^{10\varepsilon}}{(N\eta)^{\frac{5}{6}}} \ll \frac{1}{(N\eta)^{\frac{2}{3}}}$$

when ε is sufficiently small. Hence, the term $\frac{1}{N}\sum_{i=1}^N T_i \widetilde{\tau}_{i1}\Upsilon$ can be absorbed into the bound for $\mathfrak{c}_1$ in (8.30).

The second modification we need to do is the estimate for the analogue of (6.23). We take the case $j=1$ for example. In the step of (6.24), we used the estimate $\Lambda_{\mathrm{d}i}^c \prec \Psi$ from (5.16) to replace G_{ii} in the definition of ε_{i1} in (5.29) by $\frac{1}{a_i - \omega_B^c}$ in (6.24), and also the bound of T_i in (5.16) was used in (6.24). But now, lacking the conditions in (5.13), these bounds are not available. Instead, we shall need to extract similar information from the presence of the cutoff functions $\varphi(\Gamma_i)$'s and $\varphi(\Gamma)$. The analogue of ε_1 in the proof of (8.29) can be written as

$$\widetilde{\varepsilon}_1 = \frac{1}{N}\sum_{i=1}^N \varepsilon_{i1} \operatorname{tr} G \widetilde{\tau}_{i1} = \frac{1}{N}\sum_i \mathring{\boldsymbol{h}}_i^* \widetilde{B}^{\langle i\rangle} \mathring{\boldsymbol{h}}_i G_{ii} \widetilde{\tau}_{i1} + \widetilde{\delta}_1,$$

with $\widetilde{\delta}_1$ satisfying

$$\mathbb{E}|\widetilde{\delta}_1|^k = o\Big(\frac{1}{(N\eta)^{\frac{2k}{3}}}\Big) \tag{C.1}$$

for any given $k > 0$. In the estimate (C.1), we again used the fact $\frac{1}{N}\sum_{i=1}^{N}|T_i|\varphi(\Gamma) \leq C\frac{N^{2\varepsilon}}{\sqrt{N\eta}}$ to estimate the average of the second term in ε_{i1} (*cf.* (5.29)). Therefore, our task is to prove the weaker but unconditional estimate

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{N}\sum_i \mathring{\boldsymbol{h}}_i^* \widetilde{B}^{\langle i\rangle}\mathring{\boldsymbol{h}}_i G_{ii}\widetilde{\tau}_{i1}\widetilde{\mathfrak{m}}^{(p-1,p)}\Big] \\ = \mathbb{E}\big[\mathfrak{c}_{\varepsilon 1}\widetilde{\mathfrak{m}}^{(p-1,p)}\big] + \mathbb{E}\big[\mathfrak{c}_{\varepsilon 2}\widetilde{\mathfrak{m}}^{(p-2,p)}\big] + \mathbb{E}\big[\mathfrak{c}_{\varepsilon 3}\widetilde{\mathfrak{m}}^{(p-1,p-1)}\big],\end{aligned} \tag{C.2}$$

where

$$|\mathfrak{c}_{\varepsilon 1}| \leq C\hat{\Pi}, \qquad |\mathfrak{c}_{\varepsilon 2}| \leq C\hat{\Pi}^2, \qquad |\mathfrak{c}_{\varepsilon 3}| \leq C\hat{\Pi}^2, \qquad \text{on} \quad \widehat{\Omega}_2(z).$$

Moreover, the $\mathfrak{c}_{\varepsilon i}$'s also admit the moment bound $\mathbb{E}|\mathfrak{c}_{\varepsilon i}|^k = O(1)$ for any given $k > 0$. The proof of (C.2) can be done basically in the same way as the estimate for (6.25), we thus omit the details.

The last and also the major modification is: the smooth cutoffs $\varphi(\Gamma_i)$ and $\varphi(\Gamma)$ bring in new terms during the integration by parts. More specifically, we will need to consider the derivative of the cutoffs. The derivatives of the cutoffs $\varphi(\Gamma_i)$'s can be treated similarly to the case in the proof of (8.17). In the sequel, we investigate the derivative of the term $\varphi(\Gamma)$. For instance, in the analogue of the step (6.16), the counterpart of the third term on the right side of (6.16) will be

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{k}^{(i)}\mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i\frac{\partial(\operatorname{tr} G\widetilde{\tau}_{i1})}{\partial g_{ik}}\widetilde{\mathfrak{m}}^{(p-1,p)}\Big]. \tag{C.3}$$

One new term in $\frac{\partial(\operatorname{tr} G\widetilde{\tau}_{i1})}{\partial g_{ik}}$ is

$$d_i \operatorname{tr} G\, \varphi(\Gamma_i)\varphi'(\Gamma)\frac{\partial \Gamma}{\partial g_{ik}}. \tag{C.4}$$

In the sequel, we show the contribution of the term (C.4) to (C.3). The other terms involving the derivatives of the cutoffs can be treated similarly. To show the contribution of (C.4), it suffices to prove the following three estimates

$$(c\operatorname{Im} m_{\mu_A\boxplus\mu_B} + \hat{\Lambda})^{-2}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{k}^{(i)}\hat{d}_i\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i\frac{\partial(|\Lambda_A|^2+|\Lambda_B|^2)}{\partial g_{ik}} \leq C\hat{\Pi}, \tag{C.5}$$

$$\Big(\frac{N^{5\varepsilon}}{(N\eta)^{\frac{1}{3}}}\Big)^{-2}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{k}^{(i)}\hat{d}_i\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i\frac{\partial|\Upsilon|^2}{\partial g_{ik}} \leq C\hat{\Pi}, \tag{C.6}$$

$$\Big(\frac{N^{5\varepsilon}}{\sqrt{N\eta}}\Big)^{-1}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{k}^{(i)}\hat{d}_i\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i\frac{\partial\frac{1}{N}\sum_{j=1}^{N}(|T_j|^2+N^{-1})^{\frac{1}{2}}}{\partial g_{ik}} \leq C\hat{\Pi}, \tag{C.7}$$

where we introduced the shorthand notation

$$\hat{d}_i := d_i \mathrm{tr}\, G\, \varphi(\Gamma_i)\varphi'(\Gamma)\frac{1}{\|\boldsymbol{g}_i\|}.$$

To show (C.5), we first note that $|\Lambda_\iota|^2 = \Lambda_\iota\overline{\Lambda_\iota}$, and

$$|\Lambda_\iota| \le C(c\mathrm{Im}\, m_{\mu_A\boxplus\mu_B} + \hat{\Lambda}), \qquad \iota = A, B, \tag{C.8}$$

if $\hat{d}_i \ne 0$ for at least one i by the definition of Γ in (8.27) and $\varphi'(\Gamma) \ne 0$ implying $\Gamma \le C$. Further, combining (C.8) with (7.4), we also have $|m_H - m_{\mu_A\boxplus\mu_B}| \le C(c\mathrm{Im}\, m_{\mu_A\boxplus\mu_B} + \hat{\Lambda})$. Choosing c to be sufficiently small and applying the fact $|m_{\mu_A\boxplus\mu_B}| \gtrsim 1$, we get $|m_H| \gtrsim 1$ if $\hat{d}_i \ne 0$ for at least one i. In addition, we have

$$\Big|\frac{1}{N^2}\sum_{i=1}^N \sum_k^{(i)} \hat{d}_i \boldsymbol{e}_k^* \widetilde{B}^{\langle i\rangle} G \boldsymbol{e}_i \frac{\partial \Lambda_\iota}{\partial g_{ik}}\Big| \le C\Psi^2\hat{\Pi}^2, \tag{C.9}$$

which follows from the quantitative version of (6.22). The same estimate holds if we replace Λ_ι by $\overline{\Lambda_\iota}$. Then by the simple fact $\partial|\Lambda_\iota|^2/\partial g_{ik} = \overline{\Lambda_\iota}\partial\Lambda_\iota/\partial g_{ik} + \Lambda_\iota\partial\overline{\Lambda_\iota}/\partial g_{ik}$, and the estimates (C.8) and (C.9), we see that the left side of (C.5) is actually bounded by $C\Psi^4$, which is much smaller than $\hat{\Pi}$, under our choice of $\hat{\Lambda}$ in (8.28). The proof of (C.6) is similar to that of (C.5), we thus omit the details.

At the end, we prove (C.7). Recall from (4.5) the fact $\boldsymbol{h}_j = \mathrm{e}^{-\mathrm{i}\theta_j}\boldsymbol{u}_j = \mathrm{e}^{-\mathrm{i}\theta_j}U\boldsymbol{e}_j$. Hence, we have

$$|T_j|^2 = |\boldsymbol{h}_j^* G\boldsymbol{e}_j|^2 = (U^*G)_{jj}(G^*U)_{jj} = ((U^{\langle i\rangle})^* R_i G)_{jj}(G^* R_i U^{\langle i\rangle})_{jj}$$

for any i, j. Then we have

$$\frac{\partial(|T_j|^2 + N^{-1})^{\frac{1}{2}}}{\partial g_{ik}} = (|T_j|^2 + N^{-1})^{-\frac{1}{2}}\Big(\frac{\partial((U^{\langle i\rangle})^* R_i G)_{jj}}{\partial g_{ik}}(G^* R_i U^{\langle i\rangle})_{jj} + \frac{\partial(G^* R_i U^{\langle i\rangle})_{jj}}{\partial g_{ik}}((U^{\langle i\rangle})^* R_i G)_{jj}\Big)$$

Note that $|(G^* R_i U^{\langle i\rangle})_{jj}| = |T_j|$. In the sequel, we focus on the first term in the parenthesis above. The second term can be discussed similarly. From the definition, it is elementary to derive

$$\frac{\partial R_i}{\partial g_{ik}} = -c_i \boldsymbol{e}_k(\boldsymbol{e}_i + \boldsymbol{h}_i)^* + \Delta_R(i,k), \tag{C.10}$$

where c_i and $\Delta_R(i,k)$ are defined in (5.37) and (5.40), respectively. Applying (5.38) and (C.10), we have

$$\begin{aligned}
&\frac{\partial((U^{\langle i\rangle})^* R_i G)_{jj}}{\partial g_{ik}} \\
&\quad = c_i \boldsymbol{e}_j^* (U^{\langle i\rangle})^* R_i G \boldsymbol{e}_k (\boldsymbol{e}_i + \boldsymbol{h}_i)^* \widetilde{B}^{\langle i\rangle} R_i G \boldsymbol{e}_j + c_i \boldsymbol{e}_j^* (U^{\langle i\rangle})^* R_i G R_i \widetilde{B}^{\langle i\rangle} \boldsymbol{e}_k (\boldsymbol{e}_i + \boldsymbol{h}_i)^* G \boldsymbol{e}_j \\
&\qquad - c_i (U^{\langle i\rangle})^*_{jk} (\boldsymbol{e}_i + \boldsymbol{h}_i)^* G \boldsymbol{e}_j + \boldsymbol{e}_j^* (U^{\langle i\rangle})^* \Delta_R(i,k) G \boldsymbol{e}_j + \boldsymbol{e}_j^* (U^{\langle i\rangle})^* R_i \Delta_G(i,k) \boldsymbol{e}_j .
\end{aligned} \tag{C.11}$$

We take the first term on the right side of (C.11) for example. The contribution of this term to the left side of (C.7) reads

$$\Big(\frac{N^{4\varepsilon}}{\sqrt{N\eta}}\Big)^{-1} \frac{1}{N^3} \sum_{i,j=1}^{N} \sum_{k}^{(i)} \hat{c}_{ij} \hat{d}_i \boldsymbol{e}_k^* \widetilde{B}^{\langle i\rangle} G \boldsymbol{e}_i \boldsymbol{e}_j^* (U^{\langle i\rangle})^* R_i G \boldsymbol{e}_k (\boldsymbol{e}_i + \boldsymbol{h}_i)^* \widetilde{B}^{\langle i\rangle} R_i G \boldsymbol{e}_j, \tag{C.12}$$

where we introduced the shorthand notation

$$\hat{c}_{ij} := c_i (|T_j|^2 + N^{-1})^{-\frac{1}{2}} (G^* R_i U^{\langle i\rangle})_{jj}.$$

Let $I^{(i)}$ be the identity matrix with (i,i)-th entry replaced by 0 and let $C_i := \mathrm{diag}(\hat{c}_{i1}, \ldots, \hat{c}_{iN})$. We have

$$(\text{C.12}) = \Big(\frac{N^{4\varepsilon}}{\sqrt{N\eta}}\Big)^{-1} \frac{1}{N^3} \sum_{i}^{N} \hat{d}_i (\boldsymbol{e}_i + \boldsymbol{h}_i)^* \widetilde{B}^{\langle i\rangle} R_i G C_i (U^{\langle i\rangle})^* R_i G I^{(i)} \widetilde{B}^{\langle i\rangle} G \boldsymbol{e}_i.$$

Similarly to (B.17), we have

$$\big|(\boldsymbol{e}_i + \boldsymbol{h}_i)^* \widetilde{B}^{\langle i\rangle} R_i G C_i (U^{\langle i\rangle})^* R_i G I^{(i)} \widetilde{B}^{\langle i\rangle} G \boldsymbol{e}_i\big| \le C \frac{\operatorname{Im} G_{ii} + \operatorname{Im} \mathcal{G}_{ii}}{\eta^2}$$

Therefore, we have

$$\big|(\text{C.12})\big| \le \Big(\frac{N^{4\varepsilon}}{\sqrt{N\eta}}\Big)^{-1} \frac{\operatorname{Im} m_{\mu_A \boxplus \mu_B} + \hat{\Lambda}}{N^2 \eta^2} \ll \hat{\Pi}.$$

The contributions from the other terms in (C.11) can be estimated similarly. We thus omit the details.

Except for the modifications listed above, the rest of the proof of (8.29) is the same as that for (6.5).

References

[1] N.I. Akhiezer, The Classical Moment Problem and Some Related Questions in Analysis, Hafner Publishing Co., New York, 1965.
[2] O. Ajanki, L. Erdős, T. Krüger, Universality for general Wigner-type matrices, Probab. Theory Relat. Fields 169 (3–4) (2017) 667–727.
[3] Z.G. Bao, L. Erdős, K. Schnelli, Local stability of the free additive convolution, J. Funct. Anal. 271 (3) (2016) 672–719.

[4] Z.G. Bao, L. Erdős, K. Schnelli, Local law of addition of random matrices on optimal scale, Commun. Math. Phys. 349 (3) (2017) 947–990.
[5] Z.G. Bao, L. Erdős, K. Schnelli, Convergence rate for spectral distribution of addition of random matrices, Adv. Math. 319 (2017) 251–291.
[6] Z.G. Bao, L. Erdős, K. Schnelli, Local single ring theorem on optimal scale, Ann. Probab. 47 (3) (2019) 1270–1334.
[7] Z.G. Bao, L. Erdős, K. Schnelli, On the support of the free additive convolution, arXiv:1804.11199, 2018.
[8] S. Belinschi, A note on regularity for free convolutions, Ann. Inst. Henri Poincaré Probab. Stat. 42 (5) (2006) 635–648.
[9] S. Belinschi, H. Bercovici, A new approach to subordination results in free probability, J. Anal. Math. 101 (1) (2007) 357–365.
[10] H. Bercovici, D. Voiculescu, Free convolution of measures with unbounded support, Indiana Univ. Math. J. 42 (1993) 733–773.
[11] P. Biane, Process with free increments, Math. Z. 227 (1) (1998) 143–174.
[12] P. Bourgade, L. Erdős, H.-T. Yau, Edge universality of beta ensembles, Commun. Math. Phys. 332 (1) (2014) 261–354.
[13] Z. Che, B. Landon, Local spectral statistics of the addition of random matrices, Probab. Theory Relat. Fields 175 (1–2) (2019) 579–654.
[14] G.P. Chistyakov, F. Götze, The arithmetic of distributions in free probability theory, Cent. Eur. J. Math. 9 (2011) 997–1050.
[15] B. Collins, C. Male, The strong asymptotic freeness of Haar and deterministic matrices, Ann. Sci. Éc. Norm. Supér. (4) 47 (1) (2014) 147–163.
[16] P. Diaconis, M. Shahshahani, The subgroup algorithm for generating uniform random variables, Probab. Eng. Inf. Sci. 1 (01) (1987) 15–32.
[17] L. Erdős, A. Knowles, H.-T. Yau, Averaging fluctuations in resolvents of random band matrices, Ann. Henri Poincaré 14 (2013) 1837–1926.
[18] L. Erdős, A. Knowles, H.-T. Yau, J. Yin, The local semicircle law for a general class of random matrices, Electron. J. Probab. 18 (59) (2013) 1–58.
[19] A. Guionnet, M. Krishnapur, O. Zeitouni, The single ring theorem, Ann. Math. (2) 174 (2011) 1189–1217.
[20] V. Kargin, A concentration inequality and a local law for the sum of two random matrices, Probab. Theory Relat. Fields 154 (2012) 677–702.
[21] V. Kargin, Subordination for the sum of two random matrices, Ann. Probab. 43 (4) (2015) 2119–2150.
[22] J.O. Lee, K. Schnelli, Local deformed semicircle law and complete delocalization for Wigner matrices with random potential, J. Math. Phys. 54 (10) (2013) 103504.
[23] J.O. Lee, K. Schnelli, Local law and Tracy–Widom limit for sparse random matrices, Probab. Theory Relat. Fields 171 (1) (2018) 543–616.
[24] F. Mezzadri, How to generate random matrices from the classical compact groups, Not. Am. Math. Soc. 54 (5) (2007) 592–604.
[25] L. Pastur, V. Vasilchuk, On the law of addition of random matrices, Commun. Math. Phys. 214 (2) (2000) 249–286.
[26] C. Tracy, H. Widom, Level spacing distributions and the Airy kernel, Commun. Math. Phys. 159 (1994) 151–174.
[27] D. Voiculescu, Addition of certain non-commuting random variables, J. Funct. Anal. 66 (3) (1986) 323–346.
[28] D. Voiculescu, Limit laws for random matrices and free products, Invent. Math. 104 (1) (1991) 201–220.
[29] D. Voiculescu, The analogues of entropy and of Fisher's information theory in free probability theory, I, Commun. Math. Phys. 155 (1993) 71–92.