Spectral rigidity for addition of random matrices at the regular edge

Zhigang Bao∗
HKUST
mazghao@ust.hk

László Erdős†
IST Austria
lerdos@ist.ac.at

Kevin Schnelli‡
KTH Royal Institute of Technology
schnelli@kth.se

Abstract. 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.

Date: August 30, 2017

Keywords: Random matrices, local eigenvalue density, free convolution, spectral edge
AMS Subject Classification (2010): 46L54, 60B20

1. Introduction

The pioneering work of Voiculescu [27] 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 $\mu_A$ and $\mu_B$ of $A$ and $B$. The primary example for general relative position is asymptotic freeness that can be generated by conjugation via a Haar distributed unitary matrix. In fact, under some mild regularity condition on $\mu_A$ and $\mu_B$, 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 [19, 20, 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., [17, 21, 2] and references therein. In general, it implies a

∗Partially supported by Hong Kong RGC grant ECS 2630517.
†Partially supported by ERC Advanced Grant RANMAT No. 338804.
‡Partially supported by the Göran Gustafsson Foundation.
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., [25, 11, 22] 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 \( A + U B U^* \) where \( U \) is Haar unitary. We assume that the laws of \( A \) and \( B \) are close to continuous limiting profiles \( \mu_\alpha \) and \( \mu_\beta \) 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 \( A + U B U^* \) 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 \( N^{-\frac{1}{2}+\varepsilon} \) vicinity of the deterministic spectral edges. Previously, similar result was only known with \( o(1) \) precision, see [14] 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 [12].

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 [18, 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 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 constituents, 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 \( v \in \mathbb{C}^N \), we use \( \|v\| \) 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 \( g = (g_1, \ldots, g_N) \) be a real or complex Gaussian vector. We write \( g \sim N_{\mathbb{R}}(0, \sigma^2 I_N) \) if \( g_1, \ldots, g_N \) are independent and identically distributed (i.i.d.) \( N(0, \sigma^2) \) normal variables; and we write \( g \sim N_{\mathbb{C}}(0, \sigma^2 I_N) \) 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 \( C^{-1}|a| \leq |b| \leq C|a| \).

Finally, we use double brackets to denote index sets, i.e., for \( n_1, n_2 \in \mathbb{R} \), \([n_1, n_2] := [n_1, n_2] \cap \mathbb{Z} \).
2. Definition of the Model and main results

2.1. Model and assumptions. Let \( A \equiv A_N = \text{diag}(a_1, \ldots, a_N) \) and \( B \equiv B_N = \text{diag}(b_1, \ldots, b_N) \) 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 \( U(N) \), where \( U(N) \) is the \( N \)-dimensional unitary group. We study the following random Hermitian matrix

\[
H \equiv H_N := A + U BU^*.
\]  

(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 \( \mu_A, \mu_B \) and \( \mu_H \) be the empirical eigenvalue distributions of \( A, B, \) and \( H \), i.e.,

\[
\mu_A := \frac{1}{N} \sum_{i=1}^{N} \delta_{\lambda_i}, \quad \mu_B := \frac{1}{N} \sum_{i=1}^{N} \delta_{\lambda_i}, \quad \mu_H := \frac{1}{N} \sum_{i=1}^{N} \delta_{\lambda_i}.
\]

Assumption 2.2. We assume the following:

(i) Both density functions \( \rho_{\alpha} \) and \( \rho_{\beta} \) have single non-empty interval supports, \([E^\alpha_-, E^\alpha_+]\) and \([E^\beta_-, E^\beta_+]\), respectively, and \( \rho_{\alpha} \) and \( \rho_{\beta} \) are strictly positive in the interior of their supports.

(ii) In a small \( \delta \)-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

\[
C^{-1} \leq \frac{\rho_{\alpha}(x)}{(x - E^\alpha_+)^{t^\alpha}} \leq C, \quad \forall x \in [E^\alpha_-, E^\alpha_+ + \delta],
\]

\[
C^{-1} \leq \frac{\rho_{\beta}(x)}{(x - E^\beta_+)^{t^\beta}} \leq C, \quad \forall x \in [E^\beta_-, E^\beta_+ + \delta],
\]

hold for some positive constant \( C > 1 \).

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

\[
\sup_{z \in \mathbb{C}^+} |m_{\mu_{\alpha}}(z)| \leq C, \quad \sup_{z \in \mathbb{C}^+} |m_{\mu_{\beta}}(z)| \leq C,
\]

(2.2)

for some positive constant \( C \).

Assumption 2.1. We assume the following:

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

\[
d := d_L(\mu_A, \mu_\alpha) + d_L(\mu_B, \mu_\beta) \leq N^{-1+\epsilon},
\]

(2.3)

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

(v) For the lower edges, we have

\[
\inf \text{supp} \mu_A \geq E^\alpha_- - \delta, \quad \inf \text{supp} \mu_B \geq E^\beta_- - \delta,
\]

(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

\[
\sup \text{supp} \mu_A \leq C, \quad \sup \text{supp} \mu_B \leq C.
\]

(2.5)

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

Since [27], it is well known now that \( \mu_H \) can be weakly approximated by a deterministic probability measure, called the free additive convolution of \( \mu_A \) and \( \mu_B \). Here we briefly introduce some notations concerning the free additive convolution, which will be necessary to state our main results.
For a probability measure $\mu$ on $\mathbb{R}$, we denote by $F_\mu$ its negative reciprocal Stieltjes transform, i.e.,

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

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

$$\lim_{\eta \to \infty} \frac{F_\mu(i\eta)}{i\eta} = 1.$$ (2.7)

Conversely, if $F : \mathbb{C}^+ \to \mathbb{C}^+$ is an analytic function with $\lim_{\eta \to \infty} F(i\eta)/i\eta = 1$, then $F$ is the negative reciprocal Stieltjes transform of a probability measure $\mu$, 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 [8], Theorem 2.1 in [13]). Given two Borel probability measures, $\mu_1$ and $\mu_2$, on $\mathbb{R}$, there exist unique analytic functions, $\omega_1, \omega_2 : \mathbb{C}^+ \to \mathbb{C}^+$, such that,

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

$$\lim_{\eta \to \infty} \frac{\omega_1(i\eta)}{i\eta} = \lim_{\eta \to \infty} \frac{\omega_2(i\eta)}{i\eta} = 1;$$ (2.8)

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

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

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

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

is, in virtue of (2.8), the negative reciprocal Stieltjes transform of a probability measure $\mu$, called the free additive convolution of $\mu_1$ and $\mu_2$, denoted by $\mu \equiv \mu_1 \boxplus \mu_2$. The functions $\omega_1$ and $\omega_2$ 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 [28] in a generic situation and extended to full generality by Biane [10].

Choosing $(\mu_1, \mu_2) = (\mu_\alpha, \mu_\beta)$ in Proposition 2.3, we denote the associated subordination functions $\omega_1$ and $\omega_2$ by $\omega_\alpha$ and $\omega_\beta$, respectively. Analogously, for the choice $(\mu_1, \mu_2) = (\mu_A, \mu_B)$, we denote by $\omega_A$ and $\omega_B$ the associated subordination functions. With the above notations, we obtain from (2.9) and (2.10) the following subordination equations

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

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

The same system of equations hold if we replace the subscripts $(A, B)$ by $(\alpha, \beta)$.

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

$$E_- := \inf \text{supp} \mu_\alpha \boxplus \mu_\beta, \quad E_+ := \sup \text{supp} \mu_\alpha \boxplus \mu_\beta.$$ (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 (3.62).

**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

$$G(z) \equiv G_H(z) := \frac{1}{H - z}, \quad m_H(z) := \text{tr}G(z) = \frac{1}{N} \sum_{i=1}^N G_{ii}(z), \quad z \in \mathbb{C}^+.$$ (2.13)

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

$$m_H(z) = \int_R \frac{1}{x - z} \mu_H(x) = \frac{1}{N} \sum_{i=1}^N \frac{1}{\lambda_i - z}, \quad z \in \mathbb{C}^+.$$ (2.14)

We further set

$$K := \|A\| + \|B\| + 1.$$ (2.15)
Moreover, for any spectral parameter \( z = E + i\eta \in \mathbb{C}^+ \), we let
\[
\kappa \equiv \kappa(z) := \min \{|E - E_-|, |E - E_+|\},
\]
with \( E_k \) given in (2.12). We then introduce the following domain of the spectral parameter \( z \): For any \( 0 < a \leq b \) and \( 0 < \tau < \frac{E_0 - E_-}{2} \),
\[
\mathcal{D}_\tau(a, b) := \{ z = E + i\eta \in \mathbb{C}^+ : -\kappa \leq E \leq E_+ + \tau, \quad a \leq \eta \leq b \}.
\]
For any (small) positive constant \( \gamma > 0 \), we set
\[
\eta_m := N^{-1+\gamma}.
\]
Let \( \eta_M > 1 \) be some sufficiently large constant. In the rest of the paper, we will mainly work in the regime \( z \in \mathcal{D}_\tau(\eta_m, \eta_M) \) with sufficiently small constant \( \tau > 0 \). In particular, we usually have \( \eta_m \leq \eta \leq \eta_M \).

We also need the following definition on high-probability estimates from [16]. 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) \( D > 0 \),
\[
\Pr(\mathcal{X}^{(N)} > N^\epsilon \mathcal{Y}^{(N)}) \leq N^{-D},
\]
for sufficiently large \( N \geq N_0(\epsilon, D) \), and we write \( \mathcal{X} \prec \mathcal{Y} \) or \( \mathcal{X} = O_\epsilon(\mathcal{Y}) \). 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 \( \epsilon > 0 \). 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
\[
\left| \frac{1}{N} \sum_{i=1}^N d_i \left( G_{i, a} - \frac{1}{a_i - \omega_{\tau}(z)} \right) \right| \leq \frac{1}{N\eta}
\]
holds uniformly on \( \mathcal{D}_\tau(\eta_m, \eta_M) \) with \( \eta_m = N^{-1+\gamma} \) and any constant \( \eta_M > 0 \). In particular, choosing \( d_i = 1 \) for all \( i \in [1, N] \), we have the estimate
\[
|m_{\mathcal{H}}(z) - m_{\mu_A \oplus \mu_B}(z)| \leq \frac{1}{N\eta},
\]
uniformly on \( \mathcal{D}_\tau(\eta_m, \eta_M) \). Moreover, we have the improved estimate
\[
|m_{\mathcal{H}}(z) - m_{\mu_A \oplus \mu_B}(z)| \leq \frac{1}{N(\kappa + \eta)}
\]
uniformly for all \( z = E + i\eta \in \mathcal{D}_\tau(0, \eta_M) \) with \( E \leq E_- - N^{-\frac{\gamma}{2} + \epsilon} \). Here, \( \kappa = |E - E_-| \) is given in (2.14).

Let \( \gamma_j \) be the \( j \)-th \( N \)-quantile of \( \mu_A \oplus \mu_B \), i.e., \( \gamma_j \) is the smallest real number such that
\[
\mu_A \oplus \mu_B((-\infty, \gamma_j]) = \frac{j}{N}
\]
Similarly, we define \( \gamma_j^\ast \) to be the \( j \)-th \( N \)-quantile of \( \mu_A \oplus \mu_B \).

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 \( 1 \leq i \leq cN \),
\[
|\lambda_i - \gamma_i^\ast| \leq i^{-\frac{\gamma}{2}} N^{-\frac{\gamma}{2}}
\]
In fact, the same estimate also holds if \( \gamma_i^\ast \) is replaced with \( \gamma_i \).

With the following additional assumptions on the upper edges of \( \mu_A, \mu_B \) and \( \mu_A, \mu_B \), 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:

\begin{itemize}
  \item[(ii')] In a small $\delta$-neighborhood of the upper edges of their supports, the measures $\mu_\alpha$ and $\mu_\beta$ have a power law behavior, namely, there is a (large) constant $C \geq 1$ and exponents $-1 < \epsilon_+^\alpha, \epsilon_+^\beta < 1$ such that
    \[ C^{-1} \leq \frac{\rho_\alpha(x)}{(E_+^\alpha - x)^{\epsilon_+^\alpha}} \leq C, \quad \forall x \in [E_+^\alpha - \delta, E_+^\alpha], \]
    \[ C^{-1} \leq \frac{\rho_\beta(x)}{(E_+^\beta - x)^{\epsilon_+^\beta}} \leq C, \quad \forall x \in [E_+^\beta - \delta, E_+^\beta], \]

hold for some sufficiently small constant $\delta > 0$.

\begin{itemize}
  \item[(v')] For the upper edges of $\mu_A$ and $\mu_B$, we have
    \[ \sup \operatorname{supp} \mu_A \leq E_+^\alpha + \delta, \quad \sup \operatorname{supp} \mu_B \leq E_+^\beta + \delta, \]

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

\item[(vii)] The density function of $\mu_\alpha \oplus \mu_\beta$ has a single interval support, i.e.,
    \[ \operatorname{supp} \mu_\alpha \oplus \mu_\beta = [E_- , E_+]. \]
\end{itemize}

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
\begin{equation}
|\lambda_i - \gamma_i^*| \prec \min \{i^{-\frac{1}{3}}, (N - i + 1)^{-\frac{1}{3}}\} N^{-\frac{1}{3}}.
\end{equation}
The same estimate also holds if $\gamma_i^*$ is replaced with $\gamma_i$. Moreover, we have the following estimate on the convergence rate of $\mu_H$,
\begin{equation}
\sup_{x \in \mathbb{R}} |\mu_H((-\infty,x)) - \mu_A \oplus \mu_B((-\infty,x))| \sim \frac{1}{N}.
\end{equation}

We remark here that all of our results above also hold for the orthogonal setup, i.e., when $U$ is a random orthogonal matrix Haar distributed on the orthogonal group $O(N)$. 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 5-9. We first introduce
\begin{align*}
S_{AB} & \equiv S_{AB}(z) := (F_A(\omega_B(z)) - 1)(F_B(\omega_A(z)) - 1) - 1, \\
T_A & \equiv T_A(z) := \frac{1}{2} \left( F_A'(\omega_A(z))(F_B'(\omega_B(z)) - 1)^2 + F_B'(\omega_B(z))(F_A'(\omega_A(z)) - 1) \right), \\
T_B & \equiv T_B(z) := \frac{1}{2} \left( F_B'(\omega_B(z))(F_A'(\omega_A(z)) - 1)^2 + F_A'(\omega_A(z))(F_B'(\omega_B(z)) - 1) \right),
\end{align*}
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 $\mu_A$ and $\mu_B$, and where $\omega_A$ and $\omega_B$ are the subordination functions associated through (2.9). The main result in this section is the following proposition on the domain $D_\tau(\eta_m, \eta_M)$; 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:

\begin{itemize}
  \item[(i)] There exist strictly positive constants $k$ and $K$, such that
    \begin{align*}
    \min_i |a_i - \omega_B(z)| & \geq k, \\
    |\omega_A(z)| & \leq K,
    \end{align*}
  \end{itemize}

hold uniformly on $D_\tau(\eta_m, \eta_M)$.

\begin{itemize}
  \item[(ii)] For the Stieltjes transform $m_{\mu_A \oplus \mu_B}$ of $\mu_A \oplus \mu_B$, we have that
    \[ \operatorname{Im} m_{\mu_A \oplus \mu_B}(z) \sim \begin{cases} \frac{\sqrt{\kappa + \eta}}{\sqrt{\kappa + \eta}}, & \text{if } E \in \operatorname{supp} \mu_A \oplus \mu_B, \\
    \frac{\eta}{\sqrt{\kappa + \eta}}, & \text{if } E \notin \operatorname{supp} \mu_A \oplus \mu_B, \end{cases} \]

uniformly on $z = E + \eta \in D_\tau(\eta_m, \eta_M)$, with $\kappa$ given in (2.14).
(iii) For $S_{AB}$, $T_A$ and $T_B$ defined in (3.1), we have

$$S_{AB}(z) \sim \sqrt{\kappa + \eta}, \quad |T_A(z)| \leq C, \quad |T_B(z)| \leq C,$$

uniformly on $z \in D_T(\eta_m, \eta_M)$, for some constant $C$. In addition, for $z = E + i\eta \in D_T(\eta_m, \eta_M)$ with $|E - E_-| \leq \delta$ and $\eta \leq \delta$ for some sufficiently small constant $\delta > 0$, we also have

$$|T_A(z)| \geq c, \quad |T_B(z)| \geq c,$$

for some strictly positive constant $c = c(\delta)$.

(iv) For $\omega_A$, $\omega_B$ and $S_{AB}$ we have

$$|\omega'_A(z)| \leq C \frac{1}{\sqrt{\kappa + \eta}}, \quad |\omega'_B(z)| \leq C \frac{1}{\sqrt{\kappa + \eta}}, \quad |S'_{AB}(z)| \leq C \frac{1}{\sqrt{\kappa + \eta}},$$

for any $z \in D_T(\eta_m, \eta_M)$, for 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 $\mu_\alpha$ and $\mu_\beta$. 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 $\mu_A$ and $\mu_B$ 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 the $\mu_\alpha$ and $\mu_\beta$. We will always assume that $\mu_\alpha$ and $\mu_\beta$ satisfy Assumption 2.1. From Assumption 2.1 (iii) and Lemma 4.1 in [28], we know that

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

In addition, under Assumption 2.1, we see from Theorem 2.3 and Remark 2.4 in [7] 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 supp $\mu_\alpha = [E_\alpha^-, E_\alpha^+]$ and supp $\mu_\beta = [E_\beta^-, E_\beta^+]$. We introduce the spectral domain $\mathcal{E} \subset \mathbb{C}$ by setting

$$\mathcal{E} := \{ z \in \mathbb{C}^+ \cup \mathbb{R} : E_\alpha^- + E_\beta^- - 1 \leq \Re z \leq E_\alpha^- + E_\beta^- + 1, 0 \leq \Im z \leq \eta_M \}, \quad (3.9)$$

where $\eta_M > 0$ is any constant. By Lemma 3.1 in [26], we have that supp $\mu_\alpha \boxplus \mu_\beta \subset \mathcal{E} \cap \mathbb{R}$. FC

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

$$\sup_{z \in \mathcal{E}} (|\omega_\alpha(z)| + |\omega_\beta(z)|) \leq C. \quad (3.10)$$

**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

$$|\omega_\alpha(z)| > LM, \quad |\omega_\beta(z)| > L.$$  

(3.11)

Then we have from (2.9) that

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

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

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

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

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

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

hence for $L$ sufficiently large, we get a contradiction.

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

$$|\omega_\alpha(z)| > LM, \quad |\omega_\beta(z)| \leq L.$$  

(3.16)
Then we conclude from (2.9) that
\[
\frac{1}{\abs{m_{\mu_\alpha}(\omega_\beta(z))}} = \abs{\omega_\alpha(z) + \omega_\beta(z) - z} \geq \frac{LM}{2},
\] (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 \(\mu_\alpha\) is connected. Thus there is a constant \(c > 0\), depending on \(L\), such that \(\abs{m_{\mu_\alpha}(z')} \geq c\), for all \(z' \in \mathbb{C}^+\) with \(\abs{z'} \leq L\). 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
\[
\abs{\omega_\alpha(z)} \leq LM, \quad \abs{\omega_\beta(z)} \leq L,
\] (3.18)
for all \(z \in \mathcal{E}\). This completes the proof of Lemma 3.2. \(\square\)

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

Lemma 3.3. Let \(u \in \mathbb{R}\) with \(u \leq E_-\), then we have
\[
\Re \omega_\alpha(u) \leq E_-^\beta, \quad \Re \omega_\beta(u) \leq E_-^\alpha.
\] (3.19)
Moreover, \(\Re \omega_\alpha\) and \(\Re \omega_\beta\) are monotone increasing on \((-\infty, E_-)\).

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

In the second case, \(\Re m_{\mu_\beta}(\omega_\alpha(y')) > 0\), we have
\[
\Re m_{\mu_\beta}(\omega_\alpha(y')) = \int_{E_-^\beta}^{E_-^\alpha} \frac{(x - \Re \omega_\alpha(y'))d\mu_\beta(x)}{|x - \omega_\alpha(y')|^2} < 0.
\] (3.20)

However, since \(\Re m_{\mu_\beta}(\omega_\alpha(y')) = \Re m_{\mu_\alpha \Box \mu_\beta}(y')\), we get a contradiction as
\[
\Re m_{\mu_\alpha \Box \mu_\beta}(y') = \int_{y}^{\infty} \frac{d\mu_\alpha \Box \mu_\beta(x)}{x - y'} > 0,
\] (3.21)
by the definition of \(E_-\).

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

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

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

Lemma 3.4. Let \(\omega = \lambda + iv\), with \(v \geq 0\) and \(|\omega| \geq \vartheta\), for some small \(\vartheta > 0\). Let \(-1 < t < 1\). Then,
\[
\int_{0}^{\vartheta} \frac{x^t \, dx}{(x - \lambda)^2 + v^2} \sim \begin{cases} \frac{\lambda^t}{v^t}, & \text{if } \lambda > v, \\ |\omega|^{t-1} \sim \lambda^{t-1}, & \text{if } \lambda < 0, |\lambda| > v, \\ v^{t-1}, & \text{if } v > |\lambda|. \end{cases}
\] (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 \(\mu\). As \(F_\mu : \mathbb{C}^+ \to \mathbb{C}^+\) is analytic, and since \(\mu\) is a probability measure, it admits the representation
\[
F_\mu(z) - z = \Re F_\mu(i) + \int_{\mathbb{R}} \left( \frac{1}{x - z} - \frac{x}{1 + x^2} \right) \, d\tilde{\mu}(x),
\] (3.23)
for some finite Borel measure \(\tilde{\mu}\) on \(\mathbb{R}\). Note that \(\tilde{\mu}\) is in general not a probability measure. In particular, we have \(\tilde{\mu} \equiv 0\) if and only if \(\mu\) is supported at a single point. The following result about the support of the measure \(\tilde{\mu}\) associated with the measure \(\mu\) is of relevance.
Lemma 3.5. Let $\mu$ be a probability measure on $\mathbb{R}$ which is supported at more than two points, is of bounded support and satisfies $m_\mu(x) \neq 0$, for all $x \in \mathbb{R}\setminus \text{supp } \mu$. Then we have that
\[
\text{supp } \mu = \text{supp } \tilde{\mu},
\] (3.24)
where $\tilde{\mu}$ is the finite Borel measure associated with $\mu$ through (3.23).

Proof. Given any probability measure $\nu$ on $\mathbb{R}$, we first note that $x \in \mathbb{R}$ is in the support of $\nu$ if and only if its Stieltjes transform fails to be analytic in a neighborhood of $x$. For the measure $\mu$ from above, we have $m_\mu(x) \neq 0$ for all $x \in \mathbb{R}\setminus \text{supp } \mu$. Therefore, we know that $x \in \mathbb{R}$ is in the support of $\mu$ if and only if the reciprocal Stieltjes transform $F_\mu$ fails to be analytic in a neighborhood of $x$.

Since $\mu$ is supported at more than one point, we have $\tilde{\mu} \neq 0$ in (3.23). We then apply the same reasoning to conclude that $x \in \mathbb{R}$ is in the support of the measure $\tilde{\mu}$ if and only if $F_\mu$ fails to be analytic in a neighborhood of $x$. Thus (3.24) directly follows. 

Lemma 3.6. There is a constant $k_0 > 0$, such that
\[
\text{Re } \omega_\alpha(E_-) \leq E_\alpha^0 - k_0 , \quad \text{Re } \omega_\beta(E_-) \leq E_\beta^0 - k_0 .
\] (3.25)
Moreover, there exists a constant $C$, such that
\[
\text{Im } \omega_\alpha(z) + \text{Im } \omega_\beta(z) \leq \eta + C \text{Im } m_{\mu_\alpha \ast \mu_\beta}(z) ,
\] (3.26)
for all $z \in \mathcal{E}$. The constants $k_0$ and $C$ only depend on $\mu_\alpha$ and $\mu_\beta$.

Proof. Let $z \in \mathcal{E}$. Taking the imaginary part in the subordination equations (2.9) we get
\[
\frac{\text{Im } \omega_\alpha(z) + \text{Im } \omega_\beta(z) - \text{Im } z}{|\omega_\alpha(z) + \omega_\beta(z) - z|^2} = \text{Im } m_{\mu_\alpha \ast \mu_\beta}(z) .
\]
Thus we obtain
\[
\text{Im } \omega_\alpha(z) + \text{Im } \omega_\beta(z) = \text{Im } z + |\omega_\alpha(z) + \omega_\beta(z) - z|^2 \text{Im } m_{\mu_\alpha \ast \mu_\beta}(z) \leq \eta + C \text{Im } m_{\mu_\alpha \ast \mu_\beta}(z) ,
\]
where we used Lemma 3.2 to get the inequality. This proves (3.26).

We move on to prove the estimates in (3.25). Using
\[
\text{Im } m_{\mu_\alpha \ast \mu_\beta}(z) = \text{Im } \omega_\alpha(z) \int_{\mathbb{R}} \frac{d\mu_\beta(x)}{|x - \omega_\alpha(z)|^2} = \text{Im } \omega_\beta(z) \int_{\mathbb{R}} \frac{d\mu_\alpha(x)}{|x - \omega_\beta(z)|^2} ,
\] (3.27)
and (2.9), we can write
\[
\frac{\text{Im } m_{\mu_\alpha \ast \mu_\beta}(z)}{\text{Im } z} \left( \left( \int_{\mathbb{R}} \frac{d\mu_\alpha(x)}{|x - \omega_\alpha(z)|^2} \right)^{-1} + \left( \int_{\mathbb{R}} \frac{d\mu_\beta(x)}{|x - \omega_\beta(z)|^2} \right)^{-1} \right) - 1 = \frac{\text{Im } m_{\mu_\alpha \ast \mu_\beta}(z)}{\text{Im } z} \frac{1}{|m_{\mu_\alpha \ast \mu_\beta}(z)|^2} ,
\]
for all $z \in \mathcal{E} \cap \mathbb{C}^+$. Since $\text{Im } m_{\mu_\alpha \ast \mu_\beta}(z)/\text{Im } z > 0$, for all $z \in \mathcal{E} \cap \mathbb{C}^+$, we obtain
\[
\left( \int_{\mathbb{R}} \frac{d\mu_\alpha(x)}{|x - \omega_\alpha(z)|^2} \right)^{-1} + \left( \int_{\mathbb{R}} \frac{d\mu_\beta(x)}{|x - \omega_\beta(z)|^2} \right)^{-1} \geq |m_{\mu_\alpha \ast \mu_\beta}(z)|^{-2} = |\omega_\alpha(z) + \omega_\beta(z) - z|^2 ,
\] (3.28)
for all $z \in \mathcal{E} \cap \mathbb{C}^+$, and we can take the limit $\text{Im } z \to 0$ to obtain the conclusion also for $z \in \mathcal{E}$.

Next, we introduce the quantities
\[
d_\alpha := |\text{Re } \omega_\alpha(E_-) - E_\alpha^0| , \quad d_\beta := |\text{Re } \omega_\beta(E_-) - E_\beta^0| .
\] (3.29)
We now claim that $d_\alpha \geq k_0$ and $d_\beta \geq k_0$, for some constant $k_0 > 0$. Without loss of generality, we may assume that $d_\beta \geq d_\alpha$. We then proceed by distinguishing two cases: First assume that
\[
d_\alpha \leq \epsilon k , \quad d_\beta > k ,
\] (3.30)
for some small constants $k > 0$ and $\epsilon > 0$ to be chosen below.

Recalling Lemma 3.4, we note that, for fixed small $\theta > 0$,
\[
\int_{E_\alpha^0} E_\alpha^0 + \theta \frac{d\mu_\beta(x)}{|x - \omega_\alpha(z)|^2} \sim \begin{cases}
(\text{Re } \omega_\alpha(z) - E_\alpha^0)^{\theta} , & \text{if } \text{Re } \omega_\alpha(z) - E_\alpha^0 \geq \text{Im } \omega_\alpha(z) , \\
(\text{Re } \omega_\alpha(z) - E_\alpha^0)^{\theta - 1} , & \text{if } \text{Re } \omega_\alpha(z) - E_\alpha^0 \leq -\text{Im } \omega_\alpha(z) , \\
(\text{Im } \omega_\alpha(z))^{\theta - 1} , & \text{if } \text{Im } \omega_\alpha(z) > |\text{Re } \omega_\alpha(z) - E_\alpha^0| ,
\end{cases}
\] (3.31)
uniformly on the domain $\mathcal{E}$, where we have $-1 < t^\delta < 1$. (In the limit $\Im z \to 0$, the integral may be divergent, but this does not affect the following argument.) Fixing a small $\delta > 0$ and setting $z = E_- - \delta$, we obtain from all three cases in (3.31) that
\[
\left( \int_{E_-^\delta}^{E_{-}^\delta} \frac{d\mu_{\alpha}(x)}{|x - \omega_{\alpha}(E_- - \delta)|^2} \right)^{-1} \leq c(\Re \omega_{\alpha}(E_- - \delta) - E_{-}^\delta)^{1-t^\delta} \leq c(d_{\alpha})^{1-t^\delta},
\]
where we used that $\Re \omega_{\alpha}(y - \delta)$ is a non-positive increasing function as $\delta$ decreases by Lemma 3.3. In particular we can take the limit $\delta \searrow 0$.

Thus, when $d_{\alpha} < c\kappa$ and $d_\beta > k$, we have from (3.28) and (3.32) that
\[
\left( \frac{1}{\int_{E_-}^{E_{-}^\delta} \frac{d\mu_{\alpha}(x)}{|x - \omega_{\alpha}(E_- - \delta)|^2} + c(\kappa)^{1-t^\delta}} \geq |m_{\mu_{\alpha}\mu_{\beta}}(E_- - \delta)|^{-2},
\]
which implies
\[
1 \geq \int_{E_-}^{E_{-}^\delta} \frac{d\mu_{\alpha}(x)}{|x - \omega_{\beta}(E_- - \delta)|^2} \left( |m_{\mu_{\alpha}\mu_{\beta}}(E_- - \delta)|^{1-t^\delta} - c(\kappa)^{1-t^\delta} \right)
\]
\[
= \frac{1}{\int_{E_-}^{E_{-}^\delta} \frac{d\mu_{\alpha}(x)}{|x - \omega_{\beta}(E_- - \delta)|^2}} - c(\kappa)^{1-t^\delta} \int_{E_-}^{E_{-}^\delta} \frac{d\mu_{\alpha}(x)}{|x - \omega_{\beta}(E_- - \delta)|^2},
\]
where we used (2.9) to get the equality. As we are currently assuming that $d_\beta > k$, we have
\[
c(\kappa)^{1-t^\delta} \int_{E_-}^{E_{-}^\delta} \frac{d\mu_{\alpha}(x)}{|x - \omega_{\beta}(E_- - \delta)|^2} \leq c(\kappa)^{1-t^\delta} \frac{1}{d^2_\beta} \leq \epsilon c^{1-t^\delta} k^{1-t^\delta} - 1,
\]
where we used that $\Re \omega_{\alpha}(E_- - \delta)$ is a non-positive increasing function as $\delta$ decreases.

Next, as we assume that $\mu_{\beta}$ is not a single point mass, we have by the Cauchy-Schwarz inequality
\[
\int_{E_-}^{E_{-}^\delta} \frac{d\mu_{\alpha}(x)}{|x - \omega_{\beta}(E_- - \delta)|^2} \geq (1 + C_S),
\]
for some constant $C_S > 0$, uniformly for, say, all $0 \leq \delta \leq 1/10$.

Hence, returning to (3.34) and taking the limit $\delta \searrow 0$, we conclude from (3.36) and (3.37)
\[
1 \geq 1 + C_S - c\epsilon^{1-t^\delta} k^{1-t^\delta} - 1.
\]
We therefore get, for $\epsilon < (C_S/c\kappa^{1-t^\delta})^{1/(1-t^\delta)}$, for any $k > 0$, a contradiction. Here we use that $t^\delta > 1$. Thus, we can reject (3.30) for any $k$ if $\epsilon$ is sufficiently small depending on $k$.

Assume next that
\[
d_{\alpha} \leq c\kappa, \quad d_{\beta} \leq k,
\]
Following the lines from (3.31) to (3.32) with $\alpha$ and $\beta$ interchanged, we find that for any small $\delta > 0$,
\[
\left( \int_{E_-}^{E_{-}^\delta} \frac{d\mu_{\alpha}(x)}{|x - \omega_{\beta}(E_- - \delta)|^2} \right)^{-1} \leq c(\Re \omega_{\beta}(z - \delta) - E_{-}^\delta)^{1-t_{\beta}} \leq c(d_{\beta})^{1-t_{\beta}}.
\]
Hence, together with (3.32), we get from (3.28) that
\[
c(\kappa)^{1-t_{\beta}} + c\kappa^{1-t_{\beta}} \geq |m_{\mu_{\alpha}\mu_{\beta}}(E_- - \delta)|^{-2}.
\]
As $m_{\mu_{\alpha}\mu_{\beta}}(E_- - \delta)$ is increasing as $\delta$ decreases, we can take the limit $\delta \searrow 0$. Thus
\[
|m_{\mu_{\alpha}\mu_{\beta}}(E_-)|^{-2} \leq c(\kappa)^{1-t_{\beta}} + c\kappa^{1-t_{\beta}}.
\]
By (3.8), hence, since $t_{\alpha} < 1$ and $t_{\beta} < 1$, we get a contradiction by choosing $k > 0$ sufficiently small in (3.42). Thus (3.39) is ruled out. Here we only used that $\epsilon < 1$.

Combining (3.30) and (3.39), we conclude that
\[
d_{\alpha} > c\kappa, \quad d_{\beta} > k,
\]
for $\epsilon > 0$ and $k > 0$ sufficiently small. Together with (3.19) this proves (3.25) with $k_0 := c\kappa$ and concludes the proof of Lemma 3.6. □
Lemma 3.7. The lowest endpoint $E_-$ of $\text{supp} \mu_\alpha \boxplus \mu_\beta$ is the smallest real solution to the equation

\[
(F'_\mu_\alpha(\omega_\beta(z)) - 1)(F'_\mu_\beta(\omega_\alpha(z)) - 1) = 1, \quad z \in \mathbb{R}.
\]  

(3.44)

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

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

(3.45)

uniformly for all $z = E + in \in \mathcal{E}_0$ where

\[
\mathcal{E}_0 := \{ z \in \mathcal{E}_\alpha : -\kappa_0 \leq \text{Re } z - E_\alpha \leq \kappa_0, 0 \leq \text{Im } z \leq \eta_0 \}.
\]

(3.46)

Proof of Lemma 3.7. From Lemma 3.6 we know that $\text{Re } \omega_\alpha(E_-) \leq E_\alpha^2 - K$ and $\text{Re } \omega_\beta(E_-) \leq E_\alpha^2 - K$. From the subordination equations (2.9) and (3.23), we have that

\[
F'_{\mu_\beta \boxplus \mu_\alpha}(z) = F'_{\mu_\alpha}(\omega_\beta(z)) = \text{Re } F'_{\mu_\alpha}(i) + \omega_\beta(z) + \int_{\mathbb{R}} \left( \frac{1}{x - \omega_\beta(z)} - \frac{x}{1 + x^2} \right) d\tilde{\mu}_\alpha(x),
\]

(3.47)

for some Borel measures $\tilde{\mu}_\alpha$ on $\mathbb{R}$ with, according to Lemma 3.5, $\text{supp } \tilde{\mu}_\alpha = \text{supp } \mu_\alpha$. 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 $m_{\mu_\alpha \boxplus \mu_\beta}(u) = 0$. Analyticity breaks down if either $F'_{\mu_\alpha \boxplus \mu_\beta}(u) = 0$ or, according to (3.47), if $\omega_\beta(u) \in \text{supp } \mu_\alpha$, or if $\omega_\beta$ is 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 $\omega(u) \in \text{supp } \mu_\alpha$ as $\text{Re } \omega_\alpha(E_-) \leq E_\alpha^2 - k_0, k_0 > 0$. Thus $E_- \in \mathbb{R}$ is the smallest point where $\omega_\beta$ is not analytic.

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

\[
F_{\mu_\beta}(z) = F_{\mu_\alpha}(\omega_\beta(z)) = \text{Re } F_{\mu_\alpha}(i) + \omega_\beta(z) + \int_{\mathbb{R}} \left( \frac{1}{x - \omega_\beta(z)} - \frac{x}{1 + x^2} \right) d\tilde{\mu}_\beta(x),
\]

(3.48)

and $F_{\mu_\beta}$ is analytic in a disk of radius $K$ centered at $\omega = \omega_\beta(E_-)$ by (3.25). Here we also used that $\text{supp } \tilde{\mu}_\beta = \text{supp } \mu_\beta$ by Lemma 3.5. It follows that

\[
F'_{\mu_\beta}(\omega) = 1 + \int_{\mathbb{R}} \frac{d\tilde{\mu}_\beta(x)}{(x - \omega)^2},
\]

(3.49)

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

\[
\tilde{z}(\omega) := -F_{\mu_\beta}(\omega) + \omega + F'_{\mu_\beta}^{-1}(F'_{\mu_\alpha}(\omega))
\]

(3.50)

is well-defined and analytic in a neighborhood of $\omega_\beta(E_-)$. It follows from (2.9) that $\omega_\beta(z)$ is a solution $\omega = \omega_\beta(z)$ to the equation $z = \tilde{z}(\omega)$ (with $\text{Im } \omega_\beta(z) \geq \text{Im } z$). Moreover, we have $\omega_\alpha(z) = F'_{\mu_\alpha}^{-1} \circ F_{\mu_\alpha}(\omega_\beta(z))$.

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

\[
\tilde{z}(\omega) = E_- + z'(\omega_\beta(E_-))(\omega - \omega_\beta(E_-)) + \frac{1}{2} z''(\omega_\beta(E_-))(\omega - \omega_\beta(E_-))^2 + O ((\omega - \omega_\beta(E_-))^3).
\]

(3.51)

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

\[
\tilde{z}'(\omega) = -F'_{\mu_\beta}(\omega) + 1 + \frac{1}{F'_{\mu_\beta}^{-1}(F'_{\mu_\alpha} \circ F_{\mu_\alpha}(\omega))} F'_{\mu_\alpha}(\omega).
\]

(3.52)

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

\[
\tilde{z}'(\omega_\beta(z)) = -F'_{\mu_\beta}(\omega_\beta(z)) + 1 + \frac{1}{F'_{\mu_\beta}(\omega_\alpha(z))} F'_{\mu_\alpha}(\omega_\beta(z)),
\]

(3.53)

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

\[
(F'_{\mu_\alpha}(\omega_\beta(E_-)) - 1)(F'_{\mu_\beta}(\omega_\alpha(E_-)) - 1) = 1.
\]

(3.54)

This proves (3.44).
We move on to proving (3.45). From (3.50) we compute,

\[ \tilde{z}''(\omega) = -\tilde{F}_{\mu_\beta}''(\omega) + \frac{1}{\tilde{F}_{\mu_\beta}'' \circ \tilde{F}_{\mu_\beta}'} \circ \tilde{F}_{\mu_\beta}(\omega) \]

\[ \quad - \frac{1}{(\tilde{F}_{\mu_\beta}'' \circ \tilde{F}_{\mu_\beta}')(\omega))} (\tilde{F}_{\mu_\beta}'' \circ \tilde{F}_{\mu_\beta}'(\omega)) \cdot (\tilde{F}_{\mu_\beta}(\omega))^2 ; \]

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

\[ \tilde{z}''(\omega_\beta(z)) = -\tilde{F}_{\mu_\beta}''(\omega_\beta(z)) + \frac{1}{\tilde{F}_{\mu_\beta}'(\omega_\beta(z))} \tilde{F}_{\mu_\beta}''(\omega_\beta(z)) - \frac{1}{(\tilde{F}_{\mu_\beta}'(\omega_\beta(z)))^2} \tilde{F}_{\mu_\beta}(\omega_\beta(z)) \cdot (\tilde{F}_{\mu_\beta}(\omega_\beta(z)))^2 . \]

This can be re-written as

\[ \tilde{z}''(\omega_\beta(z)) = \frac{\tilde{F}_{\mu_\beta}(\omega_\beta(z))}{\tilde{F}_{\mu_\beta}'(\omega_\beta(z))} (1 - \tilde{F}_{\mu_\beta}'(\omega_\beta(z))) - \frac{1}{(\tilde{F}_{\mu_\beta}'(\omega_\beta(z)))^2} \tilde{F}_{\mu_\beta}(\omega_\beta(z)) \cdot (\tilde{F}_{\mu_\beta}(\omega_\beta(z)))^2 . \] (3.55)

Thus choosing \( z = E_- \) and recalling (3.53) and (3.54), we get

\[ \tilde{z}''(\omega_\beta(E_-)) = \frac{\tilde{F}_{\mu_\beta}(\omega_\beta(E_-))}{\tilde{F}_{\mu_\beta}'(\omega_\beta(E_-))} (1 - \tilde{F}_{\mu_\beta}'(\omega_\beta(E_-))) + \frac{\tilde{F}_{\mu_\beta}(\omega_\beta(E_-))}{\tilde{F}_{\mu_\beta}'(\omega_\beta(E_-))} (\tilde{F}_{\mu_\beta}(\omega_\beta(E_-)) - 1)^2 . \] (3.56)

From (3.49), we directly get

\[ \tilde{F}_{\mu_\beta}'(\omega_\alpha(E_-)) = 1 + \int \frac{d\tilde{\mu}_\beta(x)}{\tilde{F}_{\mu_\beta}'(\omega_\alpha(E_-)) x - \omega_\alpha(E_-)^2} , \quad \tilde{F}_{\mu_\beta}''(\omega_\beta(E_-)) = 1 + \int \frac{d\tilde{\mu}_\beta(x)}{\tilde{F}_{\mu_\beta}'(\omega_\alpha(E_-)) x - \omega_\beta(E_-)^2} . \] (3.57)

as well as

\[ \tilde{F}_{\mu_\beta}''(\omega_\alpha(E_-)) = \int \frac{d\tilde{\mu}_\beta(x)}{\tilde{F}_{\mu_\beta}'(\omega_\alpha(E_-)) x - \omega_\alpha(E_-)^2} , \quad \tilde{F}_{\mu_\beta}''(\omega_\beta(E_-)) = \frac{d\tilde{\mu}_\beta(x)}{\tilde{F}_{\mu_\beta}'(\omega_\beta(E_-)) x - \omega_\beta(E_-)^2} . \] (3.58)

Recalling that \( \omega_\alpha(E_-) \leq E_\beta^3 - K, \omega_\beta(E_-) \leq E_\beta^3 - K \) and that \( \tilde{\mu}_\alpha \neq 0, \tilde{\mu}_\beta \neq 0 \) (as \( \mu_\alpha \) and \( \mu_\beta \) are not single point masses), we infer from (3.57) and (3.58) that there are constants \( c > 0 \) and \( C < \infty \) such that

\[ c \leq \tilde{z}''(\omega_\beta(E_-)) \leq C . \] (3.59)

Choosing \( \omega = \omega_\beta(z) \) (thus \( \tilde{z}(\omega_\beta(z)) = z \)) and using \( \tilde{z}(\omega_\beta(E_-)) = 0, \tilde{z}''(\omega_\beta(E_-)) \neq 0 \) in (3.51), we get

\[ \omega_\beta(z) - \omega_\beta(E_-) = \frac{2}{\tilde{z}''(\omega_\beta(E_-))} \sqrt{E_- - z + O(|z - E_-|^3/2)} , \] (3.60)

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

Next, setting \( z = E + i\eta \), we observe that (3.59) and (3.60) imply, for \( z \) near \( E_- \), that

\[ \text{Im} \omega_\beta(z) \sim \begin{cases} \sqrt{\eta + \frac{\eta}{\sqrt{\eta}}}, & \text{if } E \geq E_- , \\ \sqrt{\eta}, & \text{if } E < E_- . \end{cases} \] (3.61)

This proves the third estimate in (3.45). The second estimate is obtained in the same way by interchanging the roles of the indices \( \alpha \) and \( \beta \). Finally the first estimate follows from (3.27) and the fact that \( \omega_\alpha(z) \) and \( \omega_\beta(z), z \in \mathcal{E}_0 \), are away from the supports of the measure \( \mu_\beta \) respectively \( \mu_\alpha \) by (3.25) and (3.60).

This shows (3.45) and concludes the proof of Lemma 3.7.

**Remark 3.8.** From (3.60) and \( m_{\mu_\alpha \oplus \mu_\beta} (z) = m_{\mu_\alpha} \omega_\beta(z) \) we get the precise behavior of \( m_{\mu_\alpha \oplus \mu_\beta} (z) \) on \( \mathcal{E}_0 \),

\[ m_{\mu_\alpha \oplus \mu_\beta} (z) m_{\mu_\alpha \oplus \mu_\beta} (E_-) = \frac{2m_{\mu_\alpha} \omega_\beta(E_-)}{\tilde{z}''(\omega_\beta(E_-))} \sqrt{E_- - z + O(|z - E_-|^3/2)} , \]

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

\[ d\mu_\alpha \oplus \mu_\beta(x) \sim \sqrt{x - E_-} dx , \quad \forall x \in [E_- E_- + \kappa_0] . \] (3.62)

**Corollary 3.9.** Let \( \mathcal{E}_0 \) be as in (3.46). Then the following behaviors hold uniformly for \( z \in \mathcal{E}_0 \),

\[ m_{\mu_\alpha \oplus \mu_\beta}'(z) \sim \frac{1}{\sqrt{|z - E_-|}} , \quad m_{\mu_\alpha \oplus \mu_\beta}''(z) \sim \frac{1}{|z - E_-|^3/2} , \] (3.63)

\[ |m_{\mu_\alpha \oplus \mu_\beta}'(z)| \sim \frac{1}{\sqrt{|z - E_-|}} , \quad |m_{\mu_\alpha \oplus \mu_\beta}''(z)| \sim \frac{1}{|z - E_-|^3/2} . \] (3.64)
and

\[ F'_{\mu_{\alpha}}(\omega_{\beta}(z)) \sim 1, \quad F'_{\mu_{\beta}}(\omega_{\beta}(z)) \sim 1, \quad F_{\mu_{\alpha}}(\omega_{\beta}(z)) \sim 1. \]  \hfill (3.65)

The same estimates hold true when the roles of the subscripts \(\alpha\) and \(\beta\) are interchanged.

**Proof.** Having established (3.45) for the behavior of \(\omega_{\alpha}\) and \(\omega_{\beta}\) around the smallest edge \(E_-\), the behaviors in (3.63) follow directly. Using the subordination equations (2.9), we note that

\[ F'_{\mu_{\alpha}}(\omega_{\beta}(z)) = F'_{\mu_{\beta}}(\omega_{\alpha}(z)) = -m_{\mu_{\alpha}} m_{\mu_{\beta}} (\omega_{\alpha}(z))^2, \]

which together with (3.63) imply (3.64). Finally, (3.65) follows directly from the analyticity of \(F_{\mu_{\beta}}\) and \(F'_{\mu_{\alpha}}\) in neighborhood of \(\omega_{\alpha}(E_-)\), respectively \(\omega_{\beta}(E_-)\). \(\square\)

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

\[ \mathcal{E}_{\kappa_0} := \{ z \in \mathcal{E} : E^\alpha_\kappa + E^\beta_\kappa - 1 \leq \text{Re} \, z - E_- \leq \kappa_0, 0 \leq \text{Im} \, z \leq \eta_M \} \]  \hfill (3.66)

with \(\kappa_0, \eta_0\) and \(\eta_M\) as in (3.46). Note that \(\mathcal{E}_0 \subset \mathcal{E}_{\kappa_0} \subset \mathcal{E}\). We further introduce the functions

\[ S_{\alpha\beta} \equiv S_{\alpha\beta}(z) := (F'_{\mu_{\alpha}}(\omega_{\beta}(z)) - 1)(F'_{\mu_{\beta}}(\omega_{\alpha}(z)) - 1) - 1, \]

\[ T_{\alpha}(z) := \frac{1}{2}(F'_{\mu_{\beta}}(\omega_{\alpha}(z))(F'_{\mu_{\beta}}(\omega_{\alpha}(z)) - 1)^2 + F_{\mu_{\beta}}(\omega_{\alpha}(z))(F_{\mu_{\beta}}(\omega_{\alpha}(z)) - 1)), \]

\[ T_{\beta}(z) := \frac{1}{2}(F'_{\mu_{\alpha}}(\omega_{\beta}(z))(F'_{\mu_{\alpha}}(\omega_{\beta}(z)) - 1)^2 + F_{\mu_{\alpha}}(\omega_{\beta}(z))(F_{\mu_{\alpha}}(\omega_{\beta}(z)) - 1)), \quad z \in \mathbb{C}^+. \]  \hfill (3.67)

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 \(m_{\mu_{\alpha}} m_{\mu_{\beta}}, \omega_{\alpha}, \omega_{\beta}\) and also the above functions.

**Corollary 3.10.** Let \(\mathcal{E}_{\kappa_0}\) be as in (3.66) and let \(\mathcal{E}_0\) be as in (3.46). Then

\[ \text{Im} \, m_{\mu_{\alpha}} m_{\mu_{\beta}}(z) \sim \text{Im} \, \omega_{\alpha}(z) \sim \text{Im} \, \omega_{\beta}(z) \sim \begin{cases} \sqrt{\kappa + \eta} & \text{if } E \geq E_-, \\ \frac{\eta}{\sqrt{\kappa + \eta}} & \text{if } E < E_- \end{cases}, \]  \hfill (3.68)

and

\[ S_{\alpha\beta}(z) \sim \sqrt{\kappa + \eta} \]  \hfill (3.69)

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

\[ T_{\alpha}(z) \sim 1, \quad T_{\beta}(z) \sim 1, \]  \hfill (3.70)

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

\[ |T_{\alpha}(z)| \leq C, \quad |T_{\beta}(z)| \leq C, \]  \hfill (3.71)

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

**Proof of Corollary 3.10.** Having established (3.45) for the behavior of \(\omega_{\alpha}\) and \(\omega_{\beta}\) on \(\mathcal{E}_0\), the behaviors in (3.68), (3.69) and (3.70) can be obtained by elementary computations using Taylor expansions as in the proof of Lemma 3.7, and the estimates in (3.57) and (3.58).

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

\[ \text{Im} \, m_{\mu_{\alpha}} m_{\mu_{\beta}}(z) = \int_{\mathbb{R}} \frac{\eta}{(x - E)^2 + \eta^2} \, d\mu_{\alpha} \, m_{\mu_{\beta}}(x) \sim \eta \]  \hfill (3.72)

uniformly on \(\mathcal{E}_{\kappa_0} \setminus \mathcal{E}_0\). Then, from (3.26), (3.72) and \(\text{Im} \, \omega_{\alpha}(z) \geq \eta, \text{Im} \, \omega_{\beta}(z) \geq \eta\), we get

\[ \text{Im} \, \omega_{\alpha}(z) \sim \eta, \quad \text{Im} \, \omega_{\beta}(z) \sim \eta. \]  \hfill (3.73)

Observe that both estimates in (3.68) are of the same order as \(\eta\) if \(z \in \mathcal{E}_{\kappa_0} \setminus \mathcal{E}_0\). Hence, we have (3.68).

Next, we show that (3.69) 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 (3.69) is comparable to 1 on \(\mathcal{E}_{\kappa_0} \setminus \mathcal{E}_0\). We first consider real \(z \in [E^\alpha_\kappa + E^\beta_\kappa - 1, E_-]\). Using (3.49) and the analogue of \(F'_{\mu_{\alpha}}\), (3.54), (3.69), the monotonicity of \(\omega_{\alpha}(z)\) and \(\omega_{\beta}(z)\) on \((-\infty, E_- - \kappa_0]\) (c.f., Lemma 3.3), and (3.25), we see that

\[ 0 \leq (F'_{\mu_{\alpha}}(\omega_{\beta}(z)) - 1)(F'_{\mu_{\beta}}(\omega_{\alpha}(z)) - 1) - 1 \leq 1 - c, \quad \forall z \in [E^\alpha_\kappa + E^\beta_\kappa - 1, E_- - \kappa_0], \]

for some small constant \(c > 0\). Hence, we have

\[ |(F'_{\mu_{\alpha}}(\omega_{\beta}(z)) - 1)(F'_{\mu_{\beta}}(\omega_{\alpha}(z)) - 1) - 1| \sim 1, \quad \forall z \in [E^\alpha_\kappa + E^\beta_\kappa - 1, E_- - \kappa_0]. \]  \hfill (3.74)
Then, (3.74) can be extended to all $z = E + i\eta$, with $E \in \left[ E^\alpha + E^\beta \,−\, 1, E_− − \kappa_0 \right]$ and $0 \leq \eta \leq \eta_0$ for sufficiently small constant $\eta_0 > 0$ by continuity. This together with (3.69) gives the estimate in the regime $E \in \left[ E^\alpha + E^\beta \,−\, 1, E_− + \kappa_0 \right]$ and $0 \leq \eta \leq \eta_0$ after possibly reducing $\eta_0$ to $\eta_0$ if $\eta_0 > \eta_0$.

It remains to show that the left side of (3.69) is proportional to 1 when $E \in \left[ E^\alpha + E^\beta \,−\, 1, E_− + \kappa_0 \right]$ and $\eta_0 \leq \eta \leq \eta_M$. To this end, we first recall (3.49), and observe from (3.47) that

$$\frac{\text{Im} F_{\muA}(\omega\beta(z)) - \text{Im} \omega\beta(z)}{\text{Im} \omega\beta(z)} = \int_{\mathbb{R}} \frac{1}{|x - \omega\beta|^\nu} \,d\muA(x).$$

Hence, using (3.49), (3.75) and their $F_{\muB}$ analogues, we have

$$|(F_{\muA}'(\omega\beta(z)) - 1)(F_{\muB}'(\omega\alpha(z)) - 1)| \leq \frac{\text{Im} F_{\muA}(\omega\beta(z)) - \text{Im} \omega\beta(z)}{\text{Im} \omega\beta(z)} \frac{\text{Im} F_{\muB}(\omega\alpha(z)) - \text{Im} \omega\alpha(z)}{\text{Im} \omega\alpha(z)}$$

$$\leq \frac{\text{Im} \omega\alpha(z) - \eta \text{Im} \omega\alpha(z) - \eta}{\text{Im} \omega\beta(z) - \text{Im} \omega\alpha(z)} \leq 1 - c,$$

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 the fact that $\eta \geq \eta_0$ and (3.73). Then, from (3.76) we get (3.69) in the whole $\mathcal{E}_{\eta_0}$.

Similarly, the upper bound in (3.71) follows from (3.73), (3.25), the monotonicity in Lemma 3.3, and the continuity of $\omega\alpha$ and $\omega\beta$. Omitting the details, we conclude the proof of Corollary 3.10.

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 $\muA$ and $\muB$ to the measures $\muA$ and $\muB$ by giving the actual proof of Proposition 3.1.

### 3.2. Proof of Proposition 3.1

In this subsection, we prove Proposition 3.1. Consider the $N$-dependent measures $\muA$ and $\muB$ while always assuming that they satisfy Assumption 2.2. Let $\omegaA(z)$ and $\omegaB(z)$ denote the subordination functions associated by (2.11) to the measures $\muA$ and $\muB$. Recall further the definition of the $z$-dependent quantities $\mathcal{S}_{\alpha \beta}$, $\mathcal{T}_A$ and $\mathcal{T}_B$ in (3.1).

Recall that $E_- = \inf \text{supp} \muA \cup \muB$. Fix sufficiently small $\varepsilon, \delta > 0$ and let the domain $\mathcal{D}$ be defined by

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

with

$$\mathcal{D}_{\text{in}} := \{z \in \mathbb{C}^+: |z - E_-| \leq \delta \} \cap \{ \text{Re} z > E_- - N^{-1+10\varepsilon}, \text{Im} z \geq N^{-1+10\varepsilon} \},$$

$$\mathcal{D}_{\text{out}} := \{z \in \mathbb{C}^+: |z - E_-| \leq \delta \} \cap \{ \text{Re} z < E_- - N^{-1+10\varepsilon} \}.$$

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 $\text{Im} z \geq N^{-1+10\varepsilon}$) unlike in case of the $\alpha, \beta$-quantities.

#### Lemma 3.11

Let $\muA, \muB, \muA$ and $\muB$ satisfy Assumptions 2.1 and 2.2. Then, there is a constant $c > 0$ such that for any $z \in \mathcal{D}$ we have

$$|\omegaA(z) - \omegaA(z)| + |\omegaB(z) - \omegaB(z)| \lesssim \frac{N^{-1+\varepsilon}}{|z - E_-|} \leq N^{-1/2+\varepsilon},$$

(3.77)

$$|\mathcal{S}_{\alpha \beta}(z)| \sim \sqrt{|z - E_-|},$$

(3.78)

and

$$|\mathcal{T}_A(z)| \sim 1, \quad |\mathcal{T}_B(z)| \sim 1,$$

(3.79)

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

$$\text{Im} m_{\muA \muB}(z) \sim \sqrt{|z - E_-|}, \quad z \in \mathcal{D}_{\text{in}},$$

(3.80)

$$\text{Im} m_{\muA \muB}(z) \lesssim \frac{\text{Im} z + O(N^{-1+\varepsilon})}{\sqrt{|z - E_-|}}, \quad z \in \mathcal{D}_{\text{out}},$$

(3.81)

for $N$ sufficiently large. Furthermore, for the imaginary parts the bound (3.77) is, for $N$ sufficiently large, sharpened to

$$|\text{Im} \omegaA - \text{Im} \omegaA| + |\text{Im} \omegaB - \text{Im} \omegaB| \leq \frac{(\text{Im} \omegaA + \text{Im} \omegaB)N^{-1+\varepsilon} + \text{Im} z}{\sqrt{|z - E_-|}},$$

(3.82)
for $z \in \mathcal{D}_{\text{out}}$, $\eta \leq N^{-1}$, which implies that

$$\inf \sup \mu_A \boxplus \mu_B \geq E_- - N^{-1+10\varepsilon}.$$  

Away from the edge we have the following weaker versions of (3.78), (3.79):

$$|S_{AB}(z)| \lesssim 1,$$  

$$|T_A(z)| + |T_B(z)| \leq C,$$  

hold uniformly for any $z$ with $\eta \leq |z - E_-| \leq C$, for $N$ sufficiently large.

Proof. First, note that we can rewrite the subordination equation for $\mu_\alpha$ and $\mu_\beta$ (c.f., (2.9) with $\mu_1 = \mu_\alpha$, $\mu_2 = \mu_\beta$) as

$$F_{\mu_A}(\omega_\beta(z)) - \omega_\alpha(z) - \omega_\beta(z) + z = r_1(z),$$  

$$F_{\mu_B}(\omega_\alpha(z)) - \omega_\alpha(z) - \omega_\beta(z) + z = r_2(z),$$  

where we introduced

$$r_1(z) := F_{\mu_A}(\omega_\beta(z)) - F_{\mu_A}(\omega_\beta(z)),$$  

$$r_2(z) := F_{\mu_B}(\omega_\alpha(z)) - F_{\mu_\beta}(\omega_\alpha(z)).$$  

By Lemma 3.6 and Lemma 3.7, we know that $\omega_\beta(z)$, $z \in \mathcal{D}$, is far away from the support of $\mu_\alpha$ and also from the support of $\mu_A$, using (2.4). Hence, using Corollary 3.9 and Lemma 3.5, we have

$$|r_1(z)| \leq C d = C N^{-1+\varepsilon},$$  

$$|r_2(z)| \leq C d = C N^{-1+\varepsilon},$$  

with $d$ given in (2.3). We rely on the following local stability result of the system (3.86).

Lemma 3.12. Fix $z_0 \in \mathcal{D}$. Assume that the functions $\omega_\alpha$, $\omega_\beta$, $r_1$, $r_2 : C^+ \rightarrow C$ satisfy (3.86) with $z = z_0$. Assume moreover that there is a function $q \equiv q(z_0)$ such that

$$|\omega_A(z_0) - \omega_\alpha(z_0)| \leq q(z_0),$$  

$$|\omega_B(z_0) - \omega_\beta(z_0)| \leq q(z_0),$$  

with $S_{\alpha\beta}(z_0) q(z_0) = o(1)$ and $S_{\alpha\beta}(z_0) q(z_0) = o(1)$, with $S_{\alpha\beta}$ given in (3.67). Then we have

$$|\omega_A(z_0) - \omega_\alpha(z_0)| + |\omega_B(z_0) - \omega_\beta(z_0)| \leq 2 \frac{|r_1(z_0)| + |r_2(z_0)|}{|S_{\alpha\beta}(z_0)|},$$  

for $N$ sufficiently large.

Proof. The proof is almost identical to the proof of Proposition 4.1 in [3]. The only difference is that, by Corollary 3.9, $F''_{\mu_A}(\omega_\beta(z))$ and $F''_{\mu_B}(\omega_\alpha(z))$ are $O(1)$ uniformly in $z \in \mathcal{D}$. Hence, in (4.11) of [3], we can stop the Taylor expansion in $\Omega_{2}(z) = \omega_\beta(z) - \omega_\beta(z)$ at second order and estimate the remainder by $O(|\Omega_{2}(z)|^2)$. This means that the factor $K/k^2$ in the subsequent formulas (4.12) and (4.13) can be replaced by a constant. Recalling that the current $S_{\alpha\beta}$ plays the role of $1/S$ in [3], we find that in the equation (4.13) we are in the linear regime provided that $S_{\alpha\beta}(z) q(z_0) \ll 1$, $S_{\alpha\beta}(z) q(z_0) \ll 1$. Following the dichotomy argument of [3], we prove Lemma 3.12. We omit the details. $\square$

Continuing the proof of Lemma 3.11, we use a continuity argument to establish (3.90) with $q(z) := N^{-1+5\varepsilon} / \sqrt{|z - E_-|}$. For $z \in \mathcal{D}$ with $\text{Im} z = \eta M$, for some fixed $\eta M = O(1)$, the local linear stability result of Lemma 4.2, of [3] shows that $|\omega_A(z) - \omega_\alpha(z)| + |\omega_B(z) - \omega_\beta(z)| \leq 2 |r_1(z)| + 2 |r_2(z)| \leq N^{-1+2\varepsilon}$, provided that $\text{Im} \omega_A(z) - \text{Im} z \geq c > 0$ and $\text{Im} \omega_B(z) - \text{Im} z \geq c > 0$. These bounds follow from the subordination equation and the representation:

$$\text{Im} \omega_A(z) - \text{Im} z = \text{Im} F_{\mu_A}(\omega_B(z)) - \text{Im} \omega_B(z) = (\text{Im} z) \int_R \frac{d\tilde{\mu}_A(x)}{|x - z|^2} \geq c' > 0$$  

if $\text{Im} z \geq \eta M$, and similarly for $\omega_B$.

Using the Lipschitz continuity of the subordination functions on $\mathcal{D}$, in particular $|\omega_A'(z)|, |\omega_B'(z)| \leq \eta^{-2}$, and similar for $\omega_\alpha$ and $\omega_\beta$, we can bootstrap (3.89) and (3.90) with $q(z) = N^{-1+5\varepsilon} / \sqrt{|z - E_-|}$, as then $q(z) S_{\alpha\beta}(z) \sim N^{-5\varepsilon}$ (since $S_{\alpha\beta}(z) \sim \sqrt{|z - E_-|}$ by (3.69)). Thus we have

$$|\omega_A(z) - \omega_\alpha(z)| + |\omega_B(z) - \omega_\beta(z)| \lesssim \frac{d}{|S_{\alpha\beta}|} \leq \frac{N^{-1+\varepsilon}}{\sqrt{|z - E_-|}} \leq N^{-1/2+\varepsilon},$$  

since for $z \in \mathcal{D}$, we have $|z - E_-| \geq N^{-1+10\varepsilon}$, i.e., $|S_{\alpha\beta}(z)| \geq N^{-1/2+5\varepsilon}$, this proves (3.77).

From this bound we can compare $S_{\alpha\beta}$ and $S_{AB}$, $T_\alpha$ and $T_A$, and $T_\beta$ and $T_B$, e.g.,

$$|S_{AB}(z) - S_{\alpha\beta}(z)| \leq \left| (F''_{\mu_A}(\omega_B(z)) - 1) (F''_{\mu_B}(\omega_A(z)) - 1) - (F''_{\mu_A}(\omega_\beta(z)) - 1) (F''_{\mu_B}(\omega_\alpha(z)) - 1) \right|$$
and we have the rigidity estimate for all quantiles,

\[(F_{\mu_A}(\omega_{\beta}(z)) - 1)(F_{\mu_B}(\omega_{\alpha}(z)) - 1) - (F_{\mu_A}(\omega_{\beta}(z)) - 1)(F_{\mu_B}(\omega_{\alpha}(z)) - 1)\]
\[\leq |\omega_A(z) - \omega_{\alpha}(z)| + |\omega_B(z) - \omega_{\beta}(z)| + d \leq N^{-1/2+\varepsilon}, \quad z \in D,
\]

(in the first estimate we used that \(F\)'s are all regular and in the second we used the same in addition to (2.35) and (2.4)). Since \(|S_{\alpha\beta}| \geq N^{-1/2+\varepsilon}\) in this regime, we immediately get (3.78). The bounds (3.79), (3.80), (3.81), (3.84) 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 (3.67) and (3.63).

The proof of (3.82) requires one more argument. Outside of the support, (3.77) is not optimal for the imaginary parts. Recall \(r_1\) and \(r_2\) from (3.87), \(z \in C^+\). Clearly
\[|\text{Im} r_1(z)| \leq C(\text{Im} \omega_{\beta}(z))N^{-1+\varepsilon}, \quad |\text{Im} r_2(z)| \leq C(\text{Im} \omega_{\alpha}(z))N^{-1+\varepsilon}, \quad z \in D,
\]
since
\[\text{Im} F_{\mu_A}(\omega_{\beta}(z)) = \frac{\text{Im} m_{\mu_A}(\omega_{\beta}(z))}{|m_{\mu_A}(\omega_{\beta}(z))|^2} = \frac{\text{Im} \omega_{\beta}(z)}{|m_{\mu_A}(\omega_{\beta}(z))|^2} \int_{R} \frac{d\mu_A(x)}{|x - \omega_{\beta}(z)|^2},
\]
so changing \(A\) to \(\alpha\) yields a factor \(N^{-1+\varepsilon}\) by (2.3) since \(\omega_{\beta}(z)\) is away from the support of \(\mu_A\). Taking imaginary parts in (3.86) and using the representations from (3.23) gives,
\[\text{Im} \omega_{\beta}(z) \int_{R} \frac{d\mu_A(x)}{|x - \omega_{\beta}(z)|^2} = \text{Im} \omega_{\alpha}(z) + \text{Im} z = \text{Im} r_1(z) = O(\text{Im} \omega_{\beta}(z)N^{-1+\varepsilon}),
\]
\[\text{Im} \omega_{\alpha}(z) \int_{R} \frac{d\mu_B(x)}{|x - \omega_{\alpha}(z)|^2} = \text{Im} \omega_{\beta}(z) + \text{Im} z = \text{Im} r_2(z) = O(\text{Im} \omega_{\alpha}(z)N^{-1+\varepsilon}),
\]

(3.91)\(z \in D\), and similarly, starting from the subordination equations for \(\mu_A\) and \(\mu_B\), we have
\[\text{Im} \omega_B(z) \int_{R} \frac{d\mu_B(x)}{|x - \omega_B(z)|^2} = \text{Im} \omega_A(z) + \text{Im} z = 0,
\]
\[\text{Im} \omega_A(z) \int_{R} \frac{d\mu_A(x)}{|x - \omega_A(z)|^2} = \text{Im} \omega_B(z) + \text{Im} z = 0.
\]

(3.92)

In fact, we can change \(\omega_{\beta}\) to \(\omega_B\) and \(\omega_{\alpha}\) to \(\omega_A\) in (3.91), to get
\[\text{Im} \omega_{\beta}(z) \int_{R} \frac{d\mu_A(x)}{|x - \omega_{\beta}(z)|^2} = \text{Im} \omega_{\alpha}(z) + \text{Im} z = O(\text{Im} \omega_{\beta}(z)N^{-1+\varepsilon}),
\]
\[\text{Im} \omega_{\alpha}(z) \int_{R} \frac{d\mu_B(x)}{|x - \omega_{\alpha}(z)|^2} = \text{Im} \omega_{\beta}(z) + \text{Im} z = O(\text{Im} \omega_{\alpha}(z)N^{-1+\varepsilon}),
\]

(3.93)\(z \in D\). Subtracting (3.92) from (3.93) and using that for very small \(\eta\) the determinant of the resulting linear system is very close to \(S_{AB}(z) \sim \sqrt{|z - E_-|}, \quad z \in D\), from (3.78), we have proved (3.82).

To prove (3.83), let \(z = x + iy\) with \(x \leq E_- - N^{-1+10\varepsilon}\). At a distance of at least \(N^{-1}\) below \(E_-\), we get
\[\text{Im} m_{\mu_A \otimes \mu_B}(z) = \text{Im} z \int_{R} \frac{d\mu_{\alpha \otimes \beta}(x)}{|x - z|^2} \leq N\text{Im} z.
\]

Moreover from \(m_{\mu_A \otimes \mu_B}(z) = m_{\alpha}(\omega_{\beta}(z))\) we have \(\text{Im} m_{\alpha}(\omega_{\beta}(z)) \sim \text{Im} \omega_{\beta}(z)\) since \(\omega_{\beta}(z)\) is away from the support of \(\mu_{\alpha}\). The same holds for \(\omega_{\alpha}(z)\), so we get \(\text{Im} \omega_{\alpha}(z) + \text{Im} \omega_{\beta}(z) \leq N\text{Im} z\). Taking \(\eta \searrow 0\), we note that the right hand side of (3.82) goes to zero. Thus we get \(\text{Im} \omega_A(z) = \text{Im} \omega_B(z) = 0\). Since \(\text{Im} m_{\mu_A \otimes \mu_B}(z) \sim \text{Im} \omega_A(z)\) in this regime, \(x\) cannot lie in the support of \(\mu_{\alpha \otimes \beta}\). This proves (3.83). \(\square\)

Recall that \(\gamma_j\) denoted the \(j\)-th \(N\)-quantiles of \(\mu_{\alpha} \otimes \mu_{\beta}\) from (2.20) and similarly let \(\gamma_j^*\) denote the \(j\)-th \(N\)-quantiles of \(\mu_{\alpha} \otimes \mu_B\), i.e., these are the smallest numbers \(\gamma_j\) and \(\gamma_j^*\) such that
\[\mu_{\alpha} \otimes \mu_B((-\infty, \gamma_j]) = \mu_{\alpha} \otimes \mu_B((-\infty, \gamma_j^*]) = \frac{j}{N}.
\]

**Lemma 3.13 (Rigidity).** Suppose Assumptions 2.1 and 2.2 hold, then we have the rigidity bound
\[|\gamma_j - \gamma_j^*| \leq j^{-1/3}N^{-\frac{5}{3}+\varepsilon}, \quad j \in [1, cN],
\]
for \(N\) sufficiently large and for some sufficiently small constant \(c > 0\).

Under the additional Assumption 2.7 we have the rigidity estimate for all quantiles, i.e.,
\[|\gamma_j - \gamma_j^*| \leq \min\{j^{-1/3}, (N + 1 - j)^{-1/3}\}N^{-\frac{5}{3}+\varepsilon}, \quad j \in [1, N].
\]
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 $E_{\kappa_0}$ from (3.46).

First, we establish that there are at most $N^{-\gamma_j}$-quantiles as well as $N^{-2/3+\varepsilon}$ vicinity of $E_- = \inf \supp \mu_\alpha \boxplus \mu_\beta$. This fact is immediate for the $\gamma_j$ quantiles since their distribution is given by the regular square root law, see (3.62). For the $\gamma_j^*$-quantiles, we know from (3.83) that $\gamma_j^* \geq E_- - N^{-1+10\varepsilon}$. We compute from (3.80)

$$\frac{j}{N} = \int_{-\infty}^{\gamma_j^*} d\mu_A \boxplus \mu_B = \int_{E_- - N^{-1+10\varepsilon}}^{\gamma_j^*} \mu_A \boxplus \mu_B(x) dx \leq C \int_{E_- - N^{-1+10\varepsilon}}^{\gamma_j^*} \Im m_{AEB}(x + iN^{-1+10\varepsilon}) dx$$

$$\leq C \int_{E_- - N^{-1+10\varepsilon}}^{\gamma_j^*} \left[ |x - E_-| + N^{-1+10\varepsilon} \right]^{1/2} dx \leq C|\gamma_j^* - E_-|^{3/2} + CN^{-1+10\varepsilon} |\gamma_j^* - E_-|,$$

which means that

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

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

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

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

$$\int_{E_- - N^{-1+10\varepsilon}}^{\gamma_j} \mu_A \boxplus \mu_B(x) dx \geq c \int_{E_- - N^{-1+10\varepsilon}}^{\gamma_j} \Im m_{AEB}(x + iN^{-1+10\varepsilon}) dx$$

if $|\gamma_j^* - E_-| \gg N^{-1+10\varepsilon}$. Using again (3.80) we get

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

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

$$|\gamma_j - \gamma_j^*| \leq |\gamma_j - E_-| + |\gamma_j^* - E_-| \leq CN^{-2/3+\varepsilon}, \text{ whenever } \gamma_j \leq E_- + N^{-2/3+\varepsilon}.$$  (3.97)

From the continuity of the free convolution (Proposition 4.13 of [9]) and the condition (2.3) we get

$$d_L(\mu_A \boxplus \mu_B, \mu_\alpha \boxplus \mu_\beta) \leq d_L(\mu_A, \mu_\alpha) + d_L(\mu_B, \mu_\beta) \leq N^{-1+\varepsilon}.$$

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 (3.62)) directly imply that

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

holds for any $x \leq E_- + \kappa_0$. Together with (3.97), this estimate immediately implies the bound (3.94).

For the proof of (3.95), 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 (3.94). Finally, (i') 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. These information together with (2.3) are sufficient to conclude that (3.98) hold uniformly for any $x \in \mathbb{R}$. The corresponding result (3.95) for the quantiles follows immediately.

\[ \square \]

Proof of Proposition 3.1. First, on the domain $D$, (i) of Proposition 3.1 follows from (3.77), (3.25), the assumption (2.4) and also the continuity of $\omega_\alpha$ and $\omega_\beta$. In the complementary domain $D - \{\eta_0, \eta_1\} \setminus D$, 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 $\omega_A$ and $\omega_B$ follow from the fact that $|m_{\mu_A \boxplus \mu_B}(z)| \geq c$, which can be derived from the rigidity (3.94) easily. 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 $\Im \omega_A(z), \Im \omega_B(z) \geq \eta$ directly. In the regime $\eta \leq \eta_0$, we use the continuity of $\omega_A$ and $\omega_B$, 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 $\omega_\beta(u)$ (c.f., (3.19)).

Similarly, on the domain $D$, Proposition 3.1 (ii) follows from (3.80) and (3.80) directly. In the complementary domain $D - \{\eta_0, \eta_1\} \setminus D$, we apply again the rigidity result (3.94) to conclude the proof.

Statement (iii) follows from (3.78), (3.79), (3.84) and (3.85).
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 \(F_A \equiv F_{\mu_A}, \quad F_B \equiv F_{\mu_B}\). Hence,
\[
\begin{pmatrix}
\omega'_A(z) \\
\omega'_B(z)
\end{pmatrix}
= S^{-1}
\begin{pmatrix}
F'_A(\omega_B(z)) - 1 \\
F'_B(\omega_A(z)) - 1
\end{pmatrix},
\]
where \(S \equiv S_{AB}\). Using (3.1) and (3.2) and (3.5), we directly get the first two estimates in (3.7).

Next, from the definition of \(S(z)\) in (3.1), we observe that
\[
|S'(z)| = \left|F'_B(\omega_A)(F'_A(\omega_B) - 1)\omega'_A(z) + F'_A(\omega_B)(F'_B(\omega_A) - 1)\omega'_B(z)\right| \leq C|S^{-1}(z)|,
\]
where in the second step we used (3.2), 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 a partial randomness decomposition of the Haar unitary matrix used in [4], which will often be used below.

Let \(u_i = (u_{i1}, \ldots, u_{iN})\) be the \(i\)-th column of \(U\). Let \(\theta_i\) be the argument of \(u_{ii}\). The following partial randomness decomposition of \(U\) is taken from [15] (see also [23]): For any \(i \in [1, N]\), we can write
\[
U = -e^{i\theta} R_i U^{(i)},
\]
where \(U^{(i)}\) is a unitary block-diagonal matrix whose \((i, i)\)-th entry equals 1, and its \((i, i)\)-minor is Haar distributed on \(U(N - 1)\). Hence, \(U^{(i)} e_i = e_i\) and \(e_i^* U^{(i)} = e_i^*\), where \(e_i\) is the \(i\)-th coordinate vector. Here \(R_i\) is a reflection matrix, defined as
\[
R_i := I - r_i r_i^*,
\]
where
\[
r_i := \sqrt{2} \frac{e_i + e^{-i\theta} u_i}{\|e_i + e^{-i\theta} u_i\|}. \tag{4.3}
\]
Using \(U^{(i)} e_i = e_i\) and (4.1), we see that
\[
\begin{aligned}
u_i &= U e_i = -e^{i\theta} R_i e_i, \\
&= -e^{i\theta} R_i e_i. \tag{4.4}
\end{aligned}
\]
Hence, \(R_i = R_i^*\) is actually the Householder reflection (up to a sign) sending \(e_i\) to \(-e^{-i\theta} u_i\). With the decomposition in (4.1), we can write
\[
H = A + \tilde{B} = A + R_i \tilde{B}^{(i)} R_i,
\]
where we introduced the notations
\[
\begin{aligned}
\tilde{B} := U B U^*, \\
\tilde{B}^{(i)} := U^{(i)} B(U^{(i)})^*.
\end{aligned}
\]
Observe that \(\tilde{B}^{(i)} e_i = b_i e_i\) and \(e_i^* \tilde{B}^{(i)} = b_i e_i^*\). Clearly, \(\tilde{B}^{(i)}\) is independent of \(u_i\).

It is known that \(u_i \in S_{\mathbb{C}}^{N-1} := \{x \in \mathbb{R}^N : \|x\| = 1\}\) is a uniformly distributed complex vector, and there exists a Gaussian vector \(\tilde{g}_i \sim \mathcal{N}(0, N^{-1} I_N)\) such that
\[
u_i = \tilde{g}_i/\|\tilde{g}_i\|. \tag{4.5}
\]
We then further introduce the notations
\[
\begin{aligned}
g_i := e^{-i\theta} \tilde{g}_i, \\
h_i := \tilde{g}_i/\|\tilde{g}_i\| = e^{-i\theta} u_i, \\
\ell_i := \sqrt{2} \frac{e_i}{\|e_i + h_i\|}.
\end{aligned}
\]
Observe that the components \(g_{ik}\) of \(g_i\) are independent. Moreover, for \(k \neq i\), \(g_{ik} \sim \mathcal{N}(0, \frac{1}{N})\) while \(g_{ii}\) is a \(\chi^2\)-distributed random variable with \(\mathbb{E} g_{ii}^2 = \frac{1}{\lambda} N\). With the above notations, we can write \(r_i\) in (4.3) as
\[
r_i = \ell_i (e_i + h_i). \tag{4.6}
\]
In addition, using (4.4) and the fact \(R_i^2 = I\), we have
\[
R_i e_i = -h_i, \quad R_i h_i = -e_i, \tag{4.7}
\]
which also imply

\[ h_i^* \tilde{B}^{(i)} R_i = -e_i^* \tilde{B} , \quad e_i^* \tilde{B}^{(i)} R_i = -b_i h^*_i = -h^*_i \tilde{B} . \]  

(4.8)

Here, in the first equality of the second equation we used that \( e_i^* \tilde{B}^{(i)} = b_i e_i \). We introduce the vectors

\[ \tilde{g}_i := g_i - g_{ii} e_i , \quad \tilde{h}_i := \frac{\tilde{g}_i}{\| \tilde{g}_i \|} , \]

where the \( \chi \)-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 \( \chi \)-distributed variable

\[ \text{In this subsection, we summarize the main route of the proof.} \]

\[ \text{We first introduce the following control parameters} \]

\[ \Psi \equiv \Psi(z) := \sqrt{\frac{1}{N \eta}} , \quad \Pi \equiv \Pi(z) := \sqrt{\frac{\text{Im} m_H}{N \eta}} . \]  

(4.9)

In [4], we investigated two main quantities:

\[ S_i \equiv S_i(z) : = h_i^* \tilde{B}^{(i)} Ge_i , \quad T_i \equiv T_i(z) : = h_i^* Ge_i . \]  

(4.10)

In particular we showed that

\[ S_i = \frac{z - \omega_B(z)}{a_i - \omega_B(z)} + O_{\prec}(\Psi) , \quad T_i = O_{\prec}(\Psi) , \]

by performing integration by parts in the \( h_i^* \) variable. Using the identity

\[ G_{ii} = \frac{1 - (\tilde{B} G)_{ii}}{a_i - z} \]

and that

\[ (\tilde{B} G)_{ii} = e_i^* R_i \tilde{B}^{(i)} R_i Ge_i = -h_i^* \tilde{B}^{(i)} R_i Ge_i = -S_i + h_i^* \tilde{B}^{(i)} r_i^* Ge_i , \]

\[ = -S_i + \ell_i^2 (h_i^* \tilde{B}^{(i)} h_i + b_i h_{ii})(G_{ii} + T_i) , \]

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 weights \( d_i \)) of the quantity

\[ Z_i := Q_i + G_{ii} \mathcal{Y} , \]

where we defined

\[ Q_i \equiv Q_i(z) := (\tilde{B} G)_{ii} tr G - G_{ii} tr \tilde{B} G , \]

(4.11)

\[ \mathcal{Y} \equiv \mathcal{Y}(z) := tr \tilde{B} G - (tr \tilde{B} G)^2 + tr G tr \tilde{B} G \tilde{B} . \]  

(4.12)

From the entry-wise laws it is clear that \( |Q_i| , |\mathcal{Y}| \prec \Psi \), 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 the most convenient object to start a high moment estimate for \( \frac{1}{N} \sum_i d_i Q_i \). The reason is that one step of integration by parts generates an additional term, \( G_{ii} \mathcal{Y} \), 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 that average is one order better, i.e.,

\[ \left| \frac{1}{N} \sum_{i=1}^N d_i Z_i \right| \prec \Psi^2 . \]  

(4.13)

Then, using (4.13) with \( d_i \equiv 1 \), we obtained \( |\mathcal{Y}| \prec \Psi^2 \). Thus \textit{a posteriori} we showed that the counter term \( G_{ii} \mathcal{Y} \) is irrelevant for estimates of order \( \Psi^2 \) 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 \( \tilde{B} G \) 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 $\sqrt{\text{Im} m}$, where we set $m = m_{\mu, \nu}$. This factor reflects that the density of states is small at the edge (at a square root edge we have $\text{Im} m(z) \sim \sqrt{\kappa + \eta}$, where $\eta = \text{Im} z$ and $\kappa$ is the distance of $\text{Re} z$ 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

$$P_i \equiv P_i(z) := (\tilde{B}G)_{ii}(G_{ii} + T_i)\Upsilon,$$

$$K_i \equiv K_i(z) := T_i + (b_i T_i + (\tilde{B}G)_{ii})\text{tr} G - (G_{ii} + T_i)\text{tr} (\tilde{B}G).$$

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,

$$K_i = (1 + b_i T_i - \text{tr} (\tilde{B}G))T_i + Q_i,$$

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 is divided into three parts. In the first part (Section 5) we obtain entry-wise bounds of the form

$$|K_i|, |Q_i|, |T_i|, |P_i| \prec \Psi,$$

as well as

$$|\Upsilon| \prec \Psi;$$

(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 $\sqrt{\text{Im} m}$. 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 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 even for $T_i$ 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 entryway bounds indicate, i.e., we expect $\frac{1}{N} \sum_i d_i Q_i \prec \Psi^2 = (N\eta)^{-1}$, and indeed this was proven in [5] in the bulk and could be extended to the edge. Due to the improvement at the edge, now we expect a bound of order $\Pi^2 \approx \text{Im} m/N\eta$, but we cannot obtain this in general. In this second part of the proof, we prove a bound of the form $\Pi \Psi \approx \sqrt{\text{Im} m/N\eta}$, which is “half-way” between the standard fluctuation averaging bound and the 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 $\Pi^2$ (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 $T$ to $T \prec \Pi \Psi$. In particular we could obtain an improved averaged bound on $P_i = Q_i + (G_{ii} + T_i) \Psi$ 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 $\Pi^2$ bound for the average of $Q_i$, but only for two very specially chosen weights, see (7.11)–(7.13). In fact, only the estimates on the “cross-term” need to be improved and the weights are chosen to achieve an additional cancellation. Nevertheless, linear combinations of $Q_i$’s with these two special sequences of weights are sufficient to invert the subordination equations and conclude that $\Lambda_i := \omega_i^* - \omega_i \prec \Psi^2$, $i = A, B$. We finally notice that

$$\frac{1}{N} \sum_{i=1}^N d_i \left( G_{ii} - \frac{1}{a_i - \omega_B^2} \right)$$

may be expressed as a linear combination of the $Q_i$, see (8.40), this quantity is already stochastically bounded by $\Pi \Psi \lesssim \Psi^2$ from the second part of the proof. Since replacing $\omega_B^* \omega_B$ with $\omega_B$ yields an error of at most $\Psi^2$, 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 determinstic $\Psi = (N \eta)^{-1/2}$ and $\Pi \approx (\Im m / N \eta)^{1/2}$ as control parameters. In fact, $\Pi$ is random, see (4.9), containing $\Im m$ which is $\Im m_{AB}$ up to a random error that itself depends on $\Lambda := |A_B| + |A_A|$. In the third part of the proof (Section 7) we obtain a self-consistent inequality for this random quantity $\Lambda$ (see (7.2)). Therefore an additional continuity argument in $\eta$ is necessary to conclude a deterministic bound on $\Lambda$.

### 5. Entry-wise Green function subordination

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

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

We define the approximate subordination functions as

$$\omega_A^*(z) := z - \frac{\tr A G(z)}{m_H(z)}, \quad \omega_B^*(z) := z - \frac{\tr B G(z)}{m_H(z)}, \quad z \in \mathbb{C}^+. \tag{5.2}$$

It will be seen that the functions $\omega_A^*$ and $\omega_B^*$ are good approximations of $\omega_A$ and $\omega_B$ 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 $\tilde{B}$, $H$, and $G(z)$, respectively,

$$\tilde{A} := U^* A U, \quad H := B + \tilde{A}, \quad G := G(z) := (H - z)^{-1}. \tag{5.3}$$

Observe that, by the cyclicity of the trace,

$$\omega_A^*(z) = z - \frac{\tr \tilde{A} G(z)}{m_H(z)}. \tag{5.4}$$

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

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

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

$$\hat{S}_i \equiv \hat{S}_i(z) := h_i^* \tilde{B}^{(i)} G e_i = S_i - h_i b_i G_{ii}, \quad \hat{T}_i \equiv \hat{T}_i(z) := h_i^* G e_i = T_i - h_i G_{ii}, \tag{5.5}$$

where the $\chi$ random variable $h_{ii}$ is kicked out.
Further, we denote (dropping the $z$-dependence from the notation for brevity)
\[ \Lambda_{d_i} := \left| G_{ii} - \frac{1}{a_i - \omega_B} \right|, \quad \Lambda_d := \max_i \Lambda_{d_i}, \quad \Lambda_T := \max_i |T_i|. \] (5.6)
We also define $\Lambda'_{d_i}$ and $\Lambda'_{d}$ analogously by replacing $\omega_B$ by $\omega_B^*$ in the definitions of $\Lambda_{d_i}$ and $\Lambda_d$, respectively.

In addition, we use the notations $\tilde{\Lambda}_{d_i}, \tilde{\Lambda}_d, \tilde{\Lambda}_T, \tilde{\Lambda}'_{d_i}, \tilde{\Lambda}'_d$ to represent their analogues, obtained by switching the rôles of $A$ and $B$, and the rôles of $U$ and $U^*$, in the definitions of $\Lambda_{d_i}, \Lambda_d, \Lambda_T, \Lambda'_{d_i}, \Lambda'_d$, e.g.,
\[ \Lambda'_{d_i} := \left| G_{ii} - \frac{1}{a_i - \omega_B} \right|, \quad \tilde{\Lambda}_d := \left| G_{ii} - \frac{1}{b_i - \omega_A} \right|. \] (5.7)

Recall $P_i, K_i,$ and $\Upsilon$ defined in (4.14), (4.15) and (4.12). We further observe the elementary identities
\[ \tilde{B}G = I - (A - z)G, \quad G \tilde{B} = I - G(A - z). \] (5.8)
Using the first identity in (5.8), we can rewrite $\Upsilon$ defined in (4.12) as
\[ \Upsilon = \text{tr\,}AG \text{ tr}\, \tilde{B}G - \text{tr\,} \text{tr}\, \tilde{B}GA = \frac{1}{N} \sum_{i=1}^{N} a_i \left( G_{ii} \text{tr}\, \tilde{B}G - (\tilde{B}G)_{ii} \text{tr}\, G \right). \] (5.9)

To ease the presentation, we further introduce the control parameter
\[ \Pi_i \equiv \Pi_i(z) := \sqrt{\frac{\text{Im}\,(G_{ii}(z) + \bar{G}_{ii}(z))}{N \eta}}, \quad i \in [1, N]. \] (5.10)
Note that since $\|H\| < \mathcal{K}$ (e.g., (2.13)), it is easy to see that $\text{Im}\, G_{ii}(z) \succ \eta$ and $\text{Im}\, G_{ii}(z) \succ \eta$ for all $z \in D_T(0, \eta_M)$, by spectral decomposition. This implies
\[ \frac{1}{\sqrt{N}} \lesssim \Pi_i(z), \quad \forall z \in D_T(0, \eta_M). \] (5.11)

In this section, we derive the following Green function subordination property.

**Proposition 5.1.** Suppose that the assumptions of Theorem 2.5 hold. Fix $z \in D_T(\eta_m, \eta_M)$. Assume that
\[ \Lambda_d(z) \prec N^{-\frac{3}{2}}, \quad \tilde{\Lambda}_d(z) \prec N^{-\frac{3}{2}}, \quad \Lambda_T(z) \prec 1, \quad \tilde{\Lambda}_T(z) \prec 1. \] (5.12)
Then we have, for all $i \in [1, N]$, that
\[ |P_i(z)| \prec \Psi(z), \quad |K_i(z)| \prec \Psi(z). \] (5.13)
In addition, we also have that
\[ |\Upsilon(z)| \prec \Psi(z) \] (5.14)
and, for all $i \in [1, N]$, that
\[ \Lambda'_{d_i}(z) \prec \Psi(z), \quad |T_i| \prec \Psi(z). \] (5.15)
The same statements hold if we switch the rôles of $A$ and $B$, and also the rôles of $U$ and $U^*$.

Before the actual proof of Proposition 5.1, we establish several bounds that follow from the assumption in (5.12). From the definitions in (5.6), the assumptions in (5.12), together with (3.2), we see that
\[ \max_{i \in [1, N]} |G_{ii}| \prec 1, \quad \max_{i \in [1, N]} |T_i| \prec 1. \] (5.16)
Analogously, we also have $\max_{i \in [1, N]} |G_{ii}| \prec 1$. Hence, under (5.12), we see that
\[ \max_{i \in [1, N]} \Pi_i(z) \prec \Psi(z). \]
Moreover, using the identities in (5.8), we also get from the first bound in (5.16) that
\[ \max_{i \in [1, N]} |(XGY)_{ii}| \prec 1, \quad X, Y = I or \tilde{B}. \] (5.17)
In addition, from (2.11) we see that
\[ \frac{1}{N} \sum_{i=1}^{N} a_i \omega_B(z) = m_{\mu_A}(\omega_B(z)) = m_{\mu_A \oplus \mu_B}(z). \] (5.18)
Then, the first bound in (5.12), together with (5.18), (5.8), (3.3) and (3.2), leads to the following estimates
\[
\begin{align*}
\text{tr} \, \tilde{G} & = m_{\mu A} \bar{\mathbb{E}}_{\mu B} + O_\prec (N^{-\frac{1}{2}}), \\
\text{tr} \, \tilde{\mathcal{B}}G & = (z - \omega_B)m_{\mu A} \bar{\mathbb{E}}_{\mu B} + O_\prec (N^{-\frac{1}{2}}), \\
\text{tr} \, \tilde{\mathcal{B}}G \mathcal{B} & = (\omega_B - z)(1 + (\omega_B - z)m_{\mu A} \bar{\mathbb{E}}_{\mu B}) + O_\prec (N^{-\frac{1}{2}}).
\end{align*}
\] (5.19)
Furthermore, by (3.2), (3.3), and (5.18), we see that all the above tracial quantities are \(O_\prec (1)\). This also implies that \(|T| \prec 1\), \(\text{c.f.}, (4.12)\). Moreover, from (5.2) and the first two equations in (5.19), we can get the following rough estimate under (5.12) and (3.2),
\[
\omega_B^2 = \omega_B + O_\prec (N^{-\frac{1}{2}}).
\] (5.20)

**Proof of Proposition 5.1.** To prove (5.13), it suffices to show the high order moment estimates
\[
\mathbb{E} \| [P_i]^{2p} \prec \Psi^{2p}, \quad \mathbb{E} \| K_i \|^2_p = \Psi^{2p},
\] (5.21)
for any fixed \(p \in \mathbb{N}\). Let us introduce the notations
\[
m^{(k,l)}_i \equiv P_i^{k \bar{T} l}, \quad n^{(k,l)}_i \equiv K_i^{k \bar{K} l}, \quad k, l \in \mathbb{N}, \quad i \in [1, N].
\] (5.22)
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
\[
|X| < \Psi^k, \quad \text{and} \quad \mathbb{E} |X|^q < \Psi^q,
\]
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 \(O_\prec (\Gamma)\) with some deterministic control parameter \(\Gamma\), we make the same convention.

With the definitions in (5.22) and the convention made above, we have the following recursive moment estimates. This type of estimates were used first in [22] 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.** For any fixed integer \(p \geq 1\) and any \(i \in [1, N]\), we have
\[
\begin{align*}
\mathbb{E}[m_i^{(p,p)}] & = \mathbb{E}[O_\prec (\Psi) m_i^{(p-1,p)}] + \mathbb{E}[O_\prec (\Psi^2) m_i^{(p-2,p)}] + \mathbb{E}[O_\prec (\Psi^2) m_i^{(p-1,p-1)}], \\
\mathbb{E}[n_i^{(p,p)}] & = \mathbb{E}[O_\prec (\Psi) n_i^{(p-1,p)}] + \mathbb{E}[O_\prec (\Psi^2) n_i^{(p-2,p)}] + \mathbb{E}[O_\prec (\Psi^2) n_i^{(p-1,p-1)}],
\end{align*}
\] (5.23) \(5.24)
where we made the convention \(m_i^{(0,0)} = n_i^{(0,0)} = 1\) and \(m_i^{(-1,1)} = n_i^{(-1,1)} = 0\) if \(p = 1\).

Although in the statements of Lemma 5.2, we use \(\Psi\), in the proof, we actually get better estimates in terms of \(\Psi^2\) instead of \(\Psi^4\) 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 \(\Psi^2\).

**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 \(\Psi\) used in the proof of Lemma 7.3 of [6], we aim here at controlling many errors in terms of \(\Pi\). 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
\[
\begin{align*}
\mathbb{E}[m_i^{(p,p)}] & = \mathbb{E} \| P_i m_i^{(p-1,p)} \| = \mathbb{E} \| (\mathcal{B}G)_{ii} \text{tr} \mathcal{G} m_i^{(p-1,p)} \| + \mathbb{E} \| (G_{ii} + T_{ii}) \text{tr} \mathcal{G} m_i^{(p-1,p)} \|,
\end{align*}
\] (5.25)
respectively,
\[
\begin{align*}
\mathbb{E}[n_i^{(p,p)}] & = \mathbb{E} \| K_i n_i^{(p-1,p)} \| = \mathbb{E} \| T_{ii} n_i^{(p-1,p)} \| + \mathbb{E} \| (b_i T_i + (\mathcal{B}G)_{ii}) \text{tr} \mathcal{G} - (G_{ii} + T_{ii}) \text{tr} \mathcal{G} n_i^{(p-1,p)} \|.
\end{align*}
\] (5.26)
Using the fact \(\mathbf{e}_i^* R_i = -h_i^* \) \(\text{c.f., (4.7)}\), we can write
\[
\begin{align*}
(\mathcal{B}G)_{ii} & = e_i^* R_i \tilde{R}^{(i)} R_i Ge_i = -h_i^* \tilde{B}^{(i)} R_i Ge_i + h_i^* \tilde{B}^{(i)} (e_i + h_i) (e_i + h_i)^* Ge_i \\
& = -S_i + \ell_i^2 (b_i h_{ii} + h_i^* \tilde{B}^{(i)} h_i) (G_{ii} + T_{ii}) = -\tilde{S}_i + \tilde{e}_i t_i,
\end{align*}
\] (5.27)
where $S_i$ and $\hat{S}_i$ are defined in (4.10) and (5.5), respectively, and
\[
\varepsilon_{i1} := \left( (\ell_i^2 - 1) h_i + \ell_i^2 \hat{B}^{(i)} h_i \right) G_{ii} + \ell_i^2 (h_i h_{ii} + h_i^* \hat{B}^{(i)} h_i) T_i. \tag{5.28}
\]
With the aid of Lemma A.1, it is elementary to check
\[
|h_{ii}| \lesssim \frac{1}{\sqrt{N}}, \quad |\ell_i^2 - 1| \lesssim \frac{1}{\sqrt{N}}, \quad |h_i^* \hat{B}^{(i)} h_i| \lesssim \frac{1}{\sqrt{N}},
\tag{5.29}
\]
where in the last inequality we also used the fact that $\text{tr} \hat{B}^{(i)} = \text{tr} B = 0$, under the convention (5.1).
Applying the bounds in (5.16) and (5.29), it is easy to see that
\[
|\varepsilon_{i1}| \lesssim \frac{1}{\sqrt{N}}. \tag{5.30}
\]
Substituting (5.27) and (5.30) into the first term on the right hand side of (5.25), we have
\[
E[(\hat{B} G)_{ii} \text{tr} G_{ii}^{(p-1,p)}] = -E[\hat{S}_i \text{tr} G_{ii}^{(p-1,p)}] + E[O_{\prec}(N^{-\frac{5}{2}}) m_{i}^{(p-1,p)}], \tag{5.31}
\]
where for the second term on the right hand side above we also used $\text{tr} G = O_{\prec}(1)$; c.f., (5.19). We recall the definition of $\hat{S}_i$ from (5.5) and rewrite
\[
\hat{S}_i = \sum_k^{(i)} \tilde{g}_i \frac{1}{\|g_i\|} e^\ast_k \hat{B}^{(i)} G e_i.
\]
Hereafter, we use the notation $\sum_k^{(i)}$ to represent the sum over $k \in \mathbb{N} \setminus \{i\}$. Thus, the first term on the right of (5.31) is of the form $E[\sum_k^{(i)} \tilde{g}_k \cdot \cdot \cdot ]$, where $\cdot \cdot \cdot$ can be regarded as a function of the $\tilde{g}_k$'s and the $g_k$'s. Recall the following integration by parts formula for complex centered Gaussian variables,
\[
\int_C g f(g, g) e^{-\frac{|g|^2}{2}} d^2 g = 2 \int_C \partial_g f(g, g) e^{-\frac{|g|^2}{2}} d^2 g,
\tag{5.32}
\]
for any differentiable function $f : C^2 \rightarrow C$. Applying (5.32) to the first term on the right of (5.31), we get
\[
E[\hat{S}_i \text{tr} G_{ii}^{(p-1,p)}] = \frac{1}{N} \sum_k^{(i)} E \left[ \frac{1}{\|g_i\|} \frac{\partial (e^\ast_k \hat{B}^{(i)} G e_i)}{\partial g_{ik}} \text{tr} G_{ii}^{(p-1,p)} \right] + \frac{1}{N} \sum_k^{(i)} E \left[ \frac{e^\ast_k \hat{B}^{(i)} G e_i}{\|g_i\|} \right] + \frac{1}{N} \sum_k^{(i)} E \left[ \frac{\partial \text{tr} G}{\partial g_{ik}} m_{i}^{(p-2, p)} \right] \tag{5.33}
\]
Analogously, by $T_i = T_i + h_{ii} G_{ii}$, (5.5), the first bound in (5.16), the first bound in (5.29), and also (5.11), we can write the first term on the right hand side of (5.26) as
\[
E[T_i n_{i}^{(p-1,p)}] = E[T_i n_{i}^{(p-1,p)}] + E[O_{\prec}(N^{-\frac{5}{2}}) n_{i}^{(p-1,p)}]. \tag{5.34}
\]
Similarly to (5.33), applying the integration by parts formula, we obtain
\[
E[T_i n_{i}^{(p-1,p)}] = \frac{1}{N} \sum_k^{(i)} E \left[ \frac{1}{\|g_i\|} \frac{\partial (e^\ast_k G e_i)}{\partial g_{ik}} n_{i}^{(p-1,p)} \right] + \frac{1}{N} \sum_k^{(i)} E \left[ \frac{\partial |g_i|^{-1}}{\partial g_{ik}} e^\ast_k G e_i n_{i}^{(p-1,p)} \right] + \frac{1}{N} \sum_k^{(i)} E \left[ \frac{\partial \text{tr} G}{\partial g_{ik}} n_{i}^{(p-2, p)} \right] \tag{5.35}
\]
First, we consider the first term on the right side of (5.33). Recall $\ell_i$ from (4.5). For brevity, we set
\[
c_i := \ell_i^2.
\tag{5.36}
\]
It is elementary to derive that
\[
\frac{\partial G}{\partial g_{ik}} = c_i (G e_i + h_i^* \hat{B}^{(i)} R_i G + G R_i \hat{B}^{(i)} e_i + h_i^* G) + \Delta G(i, k). \tag{5.37}
\]
Here $\Delta_G(i, k)$ is a small remainder, defined as
\[
\Delta_G(i, k) := -G\Delta_R(i, k)\tilde{B}^{(i)} R_i G - GR_i\tilde{B}^{(i)} \Delta_R(i, k) G,
\]

and
\[
\Delta_R(i, k) := \frac{\ell_i^2}{2\|g_i\|^2} g_i^1 (e_i h_i^* + h_i^* e_i^* + 2h_i h_i^*) - \frac{\ell_i^4}{2\|g_i\|^4} g_i^1 g_i^1 (e_i + h_i) (e_i + h_i)^*.
\]

The $\Delta_G(i, k)$'s are irrelevant error terms. We handle quantities with $\Delta_G(i, k)$ separately in Appendix B. Similarly to (7.55) of [6], using (5.37), we can get
\[
\frac{1}{N} \sum_k (\partial e_k B(i) G e_k)_i = -c_i \frac{1}{N} \sum_k e_k \tilde{B}^{(i)} G e_k(b_i T_i + (\tilde{B} G)_{ii})
\]
\[
+ c_i \frac{1}{N} \sum_k e_k \tilde{B}^{(i)} GR_i \tilde{B}^{(i)} e_k (G_{ii} + T_i) + \frac{1}{N} \sum_k e_k \tilde{B}^{(i)} \Delta_G(i, k) e_i.
\]

Note that $T_i$ naturally appears in the first term of (5.33) after integrating by parts the $\hat{S}_i$ term. This explains why we need to study the high moments of $K_i$ to get another equation. Now, we claim that
\[
\frac{1}{N} \sum_k (\partial e_k B(i) G e_k)_i = \text{tr} B G + O_{\omega}(\Pi_2^2), \quad \frac{1}{N} \sum_k e_k \tilde{B}^{(i)} GR_i \tilde{B}^{(i)} e_k = \text{tr} B G B + O_{\omega}(\Pi_1^2),
\]

where $\Pi_i$ is given in (5.10). We state the proof for the first estimate in (5.41). Note that
\[
\frac{1}{N} \sum_k e_k \tilde{B}^{(i)} G e_k = \text{tr} \tilde{B} G - \frac{1}{N} (\tilde{B}(i) G)_{ii} = \text{tr} \tilde{B} G + O_{\omega}(\frac{1}{N}),
\]

where the last step follows from the identity $(\tilde{B}(i) G)_{ii} = b_i G_{ii}$ and (5.16). Then, using that $\tilde{B}(i) = R_i \tilde{B} R_i$ and $R_i = I - r_i r_i^*$ (c.f., (4.2)), we see that
\[
\text{tr} B G - \text{tr} \tilde{B}(i) G = \text{tr} B G - \text{tr} R_i \tilde{B} R_i G = \frac{1}{N} r_i^* B G r_i + \frac{1}{N} r_i^* B G B r_i - \frac{1}{N} r_i^* B r_i r_i^* G r_i.
\]

Using (4.6), $\ell_i = 1 + O_{\omega}(\frac{1}{\sqrt{N}})$ and $\|r_i^* B\| \lesssim 1$, we get by Cauchy-Schwarz that
\[
|r_i^* B G r_i| \lesssim \left( \|G_{ii}\|^2 + \|G h_i\|^2 \right)^{\frac{1}{2}} = \left( \frac{\text{Im}(G_{ii} + G h_i)}{\eta} \right)^{\frac{1}{2}} = \left( \frac{\text{Im}(G_{ii} + G h_i)}{\eta} \right)^{\frac{1}{2}},
\]

with $G$ given in (5.3), where in the last step we used
\[
h_i^* G h_i = u_i^* G u_i = e_i^* U^* G U e_i = G_{ii}
\]

and the identities $|G|^2 = \frac{1}{\eta} \text{Im} G$ and $|G|^2 = \frac{1}{\eta} \text{Im} G$. Similarly, we have
\[
|r_i^* G r_i| \lesssim \left( \frac{\text{Im}(G_{ii} + G h_i)}{\eta} \right)^{\frac{1}{2}}, \quad |r_i^* G r_i| \lesssim \left( \frac{\text{Im}(G_{ii} + G h_i)}{\eta} \right)^{\frac{1}{2}}.
\]

Hence, we have
\[
|\text{tr} B G - \text{tr} \tilde{B}(i) G| \lesssim \frac{1}{N} \left( \frac{\text{Im}(G_{ii} + G h_i)}{\eta} \right)^{\frac{1}{2}} \lesssim \frac{\text{Im}(G_{ii} + G h_i)}{N \eta} = O_{\omega}(\Pi_1^2),
\]

where in the second step, we used the fact $\text{Im} G_{ii}, \text{Im} G h_i \gtrsim \eta$. Combining (5.42) with (5.44) we obtain the first estimate of (5.41). The second estimate in (5.41) is proved in the same way.
Using (5.26), (5.34), (5.35) and (5.46) and the estimate $\frac{c_i}{\|g_i\|} = 1 + O_\prec(\frac{1}{\sqrt{N}})$, we obtain
\[
\mathbb{E}[n_i^{(p,p)}] = \mathbb{E}[O_\prec(\Psi)\eta_i^{(p-1,p)}] + \frac{1}{N} \sum_k \mathbb{E}\left[ \frac{\partial}{\partial g_{ik}} e_k^* Ge_i n_i^{(p-1,p)} \right] + \frac{p}{N} \sum_k \mathbb{E}\left[ \frac{\partial K_i}{\|g_i\|} n_i^{(p-2,p)} \right] + \frac{p}{N} \sum_k \mathbb{E}\left[ \frac{\partial K_i}{\|g_i\|} n_i^{(p-1, p-1)} \right].
\]
(5.47)
Then, combining (5.45) with (5.46), we obtain
\[
\frac{1}{N} \sum_k \frac{\partial (e_k^* B^{(i)} Ge_i)}{\partial g_{ik}} \text{tr } G = -c_i (G_{ii} + T_i) (\text{tr } \tilde{B} G - \Psi) + \frac{1}{N} \sum_k \frac{\partial (e_k^* Ge_i)}{\partial g_{ik}} \text{tr } \tilde{B} G + O_\prec(\Pi_1^2)
\]
\[
= -c_i (G_{ii} + T_i) (\text{tr } \tilde{B} G - \Psi) + \tilde{T}_i \text{tr } \tilde{B} G + \left( \frac{1}{N} \sum_k \frac{\partial (e_k^* Ge_i)}{\partial g_{ik}} - \tilde{T}_i \right) \text{tr } \tilde{B} G + O_\prec(\Pi_1^2).
\]
(5.48)
Recall the definition of $c_i$ from (5.36). It is elementary to check that
\[
c_i = \|g_i\| - h_{ii} - (\|g_i\|^2 - 1) + O_\prec(\frac{1}{\sqrt{N}}).
\]
(5.49)
Plugging (5.49) into (5.48) and also using the second equation in (5.5), we can write
\[
\frac{1}{N} \sum_k \frac{\partial (e_k^* B^{(i)} Ge_i)}{\partial g_{ik}} \text{tr } G = -\|g_i\| (G_{ii} \text{tr } \tilde{B} G - (G_{ii} + T_i) \Psi)
\]
\[
+ \left( \frac{1}{N} \sum_k \frac{\partial (e_k^* Ge_i)}{\partial g_{ik}} - \|g_i\| \tilde{T}_i \right) \text{tr } \tilde{B} G + \epsilon_{i2} + O_\prec(\Pi_1^2),
\]
(5.50)
where $\epsilon_{i2}$ collects irrelevant terms
\[
\epsilon_{i2} := \left( \|g_i\| - c_i \right) (G_{ii} \text{tr } \tilde{B} G - (G_{ii} + T_i) \Psi) + \left( \|g_i\| \tilde{T}_i - c_i T_i \right) \text{tr } \tilde{B} G
\]
\[
= (\|g_i\|^2 - 1) G_{ii} \text{tr } \tilde{B} G - \left( h_{ii} + (\|g_i\|^2 - 1) \right) (G_{ii} + T_i) \Psi
\]
\[
+ \left( h_{ii} + (\|g_i\|^2 - 1) \right) T_i \text{tr } \tilde{B} G + O_\prec(\frac{1}{\sqrt{N}}).
\]
(5.51)
From the estimates $|h_{ii}| < \frac{1}{\sqrt{N}}, \|g_i\| = 1 + O_\prec(\frac{1}{\sqrt{N}})$, (5.16) and the observation that the tracial quantities are $O_\prec(1)$, we see that
\[
\epsilon_{i2} = O_\prec(\frac{1}{\sqrt{N}}).
\]
(5.52)
Combining (5.25), (5.27), (5.33) and (5.50), we have
\[
\mathbb{E}[m_i^{(p,p)}] = -\mathbb{E}[(\tilde{S}_i + \epsilon_{i1}) \text{tr } G m_i^{(p-1,p)}] + \mathbb{E}[(G_{ii} + T_{ii}) \Psi m_i^{(p-1,p)}]
\]
\[
= \mathbb{E}\left[ \tilde{T}_i - \frac{1}{\|g_i\|} \sum_k \frac{\partial (e_k^* Ge_i)}{\partial g_{ik}} \right] \text{tr } \tilde{B} G m_i^{(p-1,p)} - \frac{1}{N} \sum_k \mathbb{E}\left[ \frac{\partial}{\partial g_{ik}} e_k^* B^{(i)} Ge_i m_i^{(p-1,p)} \right]
\]
\[
- \frac{1}{N} \sum_k \mathbb{E}\left[ \frac{e_k^* B^{(i)} Ge_i}{\|g_i\|} \text{tr } G \frac{\partial}{\partial g_{ik}} m_i^{(p-2,p)} \right] - \frac{p}{N} \sum_k \mathbb{E}\left[ \frac{e_k^* B^{(i)} Ge_i}{\|g_i\|} \text{tr } \tilde{T}_i \frac{\partial}{\partial g_{ik}} m_i^{(p-1,p-1)} \right] + \mathbb{E}\left[ (\epsilon_{i1} \text{tr } G - \frac{1}{\|g_i\|} \epsilon_{i2} + O_\prec(\Pi_1^2)) m_i^{(p-1,p)} \right].
\]
(5.53)
For the first term on the right of (5.53), analogously to (5.35), applying (5.32) to the $\tilde{T}_i$ term, we get
\[
\mathbb{E}\left[ \left( \tilde{T}_i - \frac{1}{\|g_i\|} \sum_k \frac{\partial (e_k^* Ge_i)}{\partial g_{ik}} \right) \text{tr } \tilde{B} G m_i^{(p-1,p)} \right]
\]
\[
\begin{align*}
&= \frac{1}{N} \sum_k \mathbb{E} \left[ \frac{1}{\|g\|} \frac{\partial}{\partial g_{ik}} \tilde{B} \frac{1}{\|g\|} e_i^* G_{e_i} \text{tr} \tilde{B} G_{m_i}^{(p-1,p)} \right] + \frac{1}{N} \sum_k \mathbb{E} \left[ \frac{\partial}{\partial g_{ik}} e_i^* G_{e_i} \text{tr} \tilde{B} G_{m_i}^{(p-1,p)} \right] \\
&\quad + \frac{p - 1}{N} \sum_k \mathbb{E}\left[ \frac{\partial}{\partial g_{ik}} \tilde{B} G_{m_i}^{(p-2,p)} \right] + \frac{p}{N} \sum_k \mathbb{E}\left[ \frac{\partial}{\partial g_{ik}} \tilde{B} G_{m_i}^{(p-1,p-1)} \right].
\end{align*}
\]  

(5.54)

Recall the estimates of \(\varepsilon_{1,1}\) and \(\varepsilon_{1,2}\) in (5.30) and (5.52), respectively, which implies that \(|\varepsilon_{1,1}| \leq \Psi\) and \(|\varepsilon_{1,2}| \leq \Psi\). Therefore, to show (5.23), it suffices to estimate the second to the fifth terms on the right side of (5.53), and all the terms on the right side of (5.54). Similarly, in light of (5.26), (5.34), and (5.46), to show (5.24), it suffices to estimate the last three terms on the right side of (5.47). 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 \(\tilde{B}^{(i)}\). Let \(Q\) be any (possibly random) diagonal matrix satisfying \(\|Q\| \leq 1\) and \(X = I\) or \(A\). We have the following estimates

\[
\begin{align*}
\frac{1}{N} \sum_k \left( \frac{\partial}{\partial g_{ik}} e_i^* X_i G_{e_i} = O_{\prec}(\frac{1}{N}), \right. \\
\frac{1}{N} \sum_k \left( \frac{\partial}{\partial g_{ik}} e_i^* X_i G_{e_i} = O_{\prec}(\Pi_i^3), \right. \\
\frac{1}{N} \sum_k \left( \frac{\partial}{\partial g_{ik}} e_i^* X_i G_{e_i} = O_{\prec}(\Pi_i^3). \right. \\
\frac{1}{N} \sum_k \left( \frac{\partial}{\partial g_{ik}} e_i^* X_i G_{e_i} = O_{\prec}(\Psi^2 \Pi_i^3). \right.
\end{align*}
\]  

(5.55)

In addition, the same estimates hold if we replace \(\frac{\partial}{\partial g_{ik}}\) and \(\frac{\partial}{\partial g_{ik}}\) by their complex conjugates \(\frac{\partial}{\partial g_{ik}}\) and \(\frac{\partial}{\partial g_{ik}}\) in the last four equations above.

The proof of Lemma 5.3 will be postponed to Appendix B. 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 (5.47), the second term of (5.53), and the second term of (5.54) follow from the first equation in (5.55). The third term of (5.53) and the first term of (5.54) can be estimated by the last equation in (5.55), after writing \(\text{tr} \tilde{B} G = 1 - \text{tr} (A - z) G\). All the other terms have \(\frac{\partial}{\partial g_{ik}}\) and \(\frac{\partial}{\partial g_{ik}}\) or their complex conjugate involved. Recall the definitions in (4.14) and (4.15), and also the first equation in (5.8). Then, by the chain rule, we see that all terms in (5.47), (5.53) and (5.54), with \(\frac{\partial}{\partial g_{ik}}\) and \(\frac{\partial}{\partial g_{ik}}\) or their complex conjugate counterparts involved, can be estimated by combining the last three equations in (5.55). This completes the proof of Lemma 5.2.

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.

First, using Young’s inequality, we obtain from (5.23) that for any given (small) \(\varepsilon > 0\),

\[
\mathbb{E}[m_{i}^{(p,p)}] \leq \frac{1}{3} \frac{1}{2^p} N^{2p+\varepsilon} \Psi^{2p} + 3 \frac{2p - 1}{2^p} N^{-\frac{2p}{2p}} \mathbb{E}[m_{i}^{(p,p)}].
\]

Since \(\varepsilon > 0\) was arbitrary, this implies the first bound in (5.21). The second one then follows from (5.24) in the same manner. By Markov’s inequality, we get (5.13).

Next, we show how (5.14) and (5.15) follow from (5.13) and the assumption (5.12). To this end, we first prove the following crude bound

\[
\Lambda_f(z) \leq N^{-\frac{3}{2}}.
\]

(5.56)

From the definition in (4.15), we can rewrite the second estimate in (5.13) as

\[
(1 + b_i \text{tr} G - \text{tr} (\tilde{B} G)) T_i = G_{ii} \text{tr} (\tilde{B} G) - (\tilde{B} G)_{ii} \text{tr} G + O_{\prec}(\Psi).
\]

(5.57)

Using the identity

\[
(\tilde{B} G)_{ii} = 1 - (a_i - z) G_{ii}(z),
\]

(5.58)
and approximate $G_{ii}$ by $(a_i - \omega_B)^{-1}$, we get from (5.12) and (3.2) that
\[(\bar{BG})_{ii} = \frac{z - \omega_B}{a_i - \omega_B} + O_\prec(N^{-\tau}).\] (5.59)

We also recall the estimates of the tracial quantities in (5.19) under the assumption (5.12). Plugging (5.59), (5.19) and the first bound in the assumption (5.12) into (5.57), we get
\[(1 + (b_i - z + \omega_B)m_{\mu_A \cap \mu_B} + O_\prec(N^{-\tau}))T_i = O_\prec(N^{-\tau}) + O_\prec(\Psi) = O_\prec(N^{-\tau}),\] (5.60)

where in the last step we used that $\Psi \leq N^{-\tau}$ for all $\eta \geq \eta_0$. From the second line in (2.11), we note that
\[1 + (b_i - z + \omega_B)m_{\mu_A \cap \mu_B} = m_{\mu_A \cap \mu_B} \left(\frac{1}{m_{\mu_A \cap \mu_B}} + b_i - z + \omega_B\right) = m_{\mu_A \cap \mu_B}(b_i - \omega_A).\]

Using (3.2) and $\|A\|, \|B\| \leq C$, we get $|m_{\mu_A \cap \mu_B}(b_i - \omega_A)| \gtrsim 1$. This together with (5.60) implies (5.56).

To prove (5.14), we recall the definition of $P_i$ in (4.14), which implies that
\[\frac{1}{N} \sum_{i=1}^N (G_{ii} + T_i) = \frac{1}{N} \sum_{i=1}^N P_i = O_\prec(\Psi).\] (5.61)

Using the facts $\frac{1}{N} \sum_{i=1}^N G_{ii} = m_{\mu_A \cap \mu_B} + O_\prec(N^{-\tau})$ (c.f., (5.19)), and $\frac{1}{N} \sum_{i=1}^N T_i = O_\prec(N^{-\tau})$, and also $|m_{\mu_A \cap \mu_B}| \gtrsim 1$, we get (5.14) from (5.61).

Then, combining (5.14) with the first estimate in (5.13), we get
\[\bar{BG}_{ii} \text{tr } G - G_{ii} \text{tr } \bar{BG} = O_\prec(\Psi).\] (5.62)

Applying the identity (5.58) and the definition of $\omega_{\bar{B}}$, we can rewrite (5.62) as
\[(a_i - \omega_{\bar{B}})G_{ii} - 1) \text{tr } G = O_\prec(\Psi).\]

As shown above that $|\text{tr } G| \gtrsim 1$ with high probability under the assumption (5.12), we get $(a_i - \omega_{\bar{B}})G_{ii} - 1 = O_\prec(\Psi)$. By (5.20) and (3.2), we also note that $|a_i - \omega_{\bar{B}}| \gtrsim 1$ with high probability. This further implies the first estimate in (5.15).

Finally, plugging (5.62) back to (5.57), we can improve the right hand side of (5.60) to $O_\prec(\Psi)$. Then the second estimate in (5.15) follows. This completes the proof of Proposition 5.1. \(\Box\)

6. Rough fluctuation averaging for general linear combinations

In this section, we prove a rough fluctuation averaging estimate for the basic quantities $Q_i$’s defined in (4.11). From (5.62), we see that
\[|Q_i| \prec \Psi.\] (6.1)

Recall the definition of the control parameters $\Pi$ and $\Pi_i$ in (4.9) and (5.10), 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 $z \in D_+(\eta_0, \eta_B)$. Suppose that the assumptions of Proposition 5.1 hold. Set $X_i = 1$ or $\bar{B}(i)$. Let $d_1, \ldots, d_N \in \mathbb{C}$ be possibly $H$-dependent quantities satisfying $\max_j |d_j| < 1$. Assume that they depend only weakly on the randomness in the sense that the following hold, for all $i, j \in [1, N]$,
\[\frac{1}{N^2} \sum_{i=1}^N \sum_{k} (i) \frac{\partial e_{ik}}{\partial q_{ik}} X_i G e_{i} = O_\prec(\Psi^2 \Pi_i^2), \quad \frac{1}{N^2} \sum_{i=1}^N \sum_{k} (i) \frac{\partial d_{ik}}{\partial q_{ik}} X_i g_k = O_\prec(\Psi^2 \Pi_i^2),\] (6.2)

and the same bounds hold when the $d_j$’s are replaced by their complex conjugates $\bar{d}_j$. Suppose that $\Pi(z) \prec \bar{\Pi}(z)$ for some deterministic and positive function $\bar{\Pi}(z)$ that satisfies $\frac{1}{\sqrt{N^2}} \prec \bar{\Pi} \prec \Psi$. Then,
\[\frac{1}{N} \sum_{i=1}^N |d_i Q_i| \prec \Psi \bar{\Pi}.\] (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
\[m^{(k, l)} := \left(\frac{1}{N} \sum_{i=1}^N d_i Q_i\right)^k \left(\frac{1}{N} \sum_{i=1}^N d_i^\tau Q_i\right)^l, \quad k, l \in \mathbb{N}.\] (6.4)
Similarly to Lemma 5.2, it suffices to prove the following recursive moment estimate.

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

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

(6.5)

**Proof of Proposition 6.1.** Similarly to the proof of (5.13) 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. \(\square\)

**Proof of Lemma 6.2.** We first claim that it suffices to prove the following statements: If $|\Upsilon(z)| < \tilde{\Upsilon}(z)$ for any deterministic and positive function $\tilde{\Upsilon}(z)$, then

$$
\mathbb{E}[m^{(p,p)}] = \mathbb{E}[O_\prec(\tilde{\Pi}^2) + O_\prec(\Psi\tilde{\Pi})m^{(p-1,p)}] + \mathbb{E}[O_\prec(\Psi^2\tilde{\Pi}^2)m^{(p-2,p)}]
$$

$$
+ \mathbb{E}[O_\prec(\Psi^2\tilde{\Pi}^2)m^{(p-1,p-1)}].
$$

(6.6)

Indeed, similarly to the proof of (5.13) from Lemma 5.2, we can again apply Young’s inequality and Markov’s inequality to get, for any $d_i$ satisfying the assumptions in Proposition 6.1, that (6.6) implies

$$
\left| \frac{1}{N} \sum_{i=1}^{N} d_i Q_i \right| < \tilde{\Pi}^2 + \Psi\tilde{\Pi} + \Psi\tilde{\Pi} < \Psi\tilde{\Pi}.
$$

(6.7)

where in the last step we used the assumption $\tilde{\Pi} < \Psi$.

Next, recall from (5.9) 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)

$$
|\Upsilon| < \Psi\tilde{\Pi} + \Psi\tilde{\Pi} < N^{-\frac{1}{2}} \tilde{\Pi}^2 + \Psi\tilde{\Pi}.
$$

(6.8)

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

$$
|\Upsilon| < \Psi\tilde{\Pi}.
$$

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

$$
\mathbb{E}[m^{(p,p)}] = \mathbb{E}[(O_\prec(\tilde{\Pi}^2) + O_\prec(\Psi^2\tilde{\Pi}))m^{(p-1,p)}] + \mathbb{E}[O_\prec(\Psi^2\tilde{\Pi}^2)m^{(p-2,p)}]
$$

$$
+ \mathbb{E}[O_\prec(\Psi^2\tilde{\Pi}^2)m^{(p-1,p-1)}].
$$

(6.9)

Observe that by the assumption that $\frac{1}{N} m \prec \tilde{\Pi}$, we also have $\Pi^2 \prec \tilde{\Pi}$ on $\mathcal{D}_\tau(\eta_{m}, \eta_{hd})$. Then the $O_\prec(\Psi^2\tilde{\Pi}^2)$ term can be absorbed by the $O_\prec(\Pi^2)$ in (6.9). Hence, we conclude (6.5) from (6.6). Therefore, in the sequel, we will focus on proving (6.6).

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

$$
\frac{1}{N} \sum_{i=1}^{N} d_i Q_i = \frac{1}{N} \sum_{i=1}^{N} (\tilde{B}G)_{ii} (d_i \text{tr } G - \frac{1}{\text{tr } G}) = \frac{1}{N} \sum_{i=1}^{N} (\tilde{B}G)_{ii} \text{tr } G \tau_{i1},
$$

(6.10)

where we introduced the notation

$$
\tau_{i1} := d_i - \frac{\frac{1}{\text{tr } G}}{\text{tr } G}.
$$

(6.11)

Similarly to the proof of (5.13), we approximate $(\tilde{B}G)_{ii}$ by $\tilde{S}_i$ (c.f., (5.27)), and then perform integration by parts using (5.32) with respect to $g_i$ in $\tilde{S}_i$. More specifically, we write

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

$$
= -\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[\tilde{S}_i \text{tr } G \tau_{i1} m^{(p-1,p)}] + \mathbb{E}[\xi_1 m^{(p-1,p)}],
$$

(6.12)
where we used the notation
\[
\varepsilon_1 := \frac{1}{N} \sum_{i=1}^{N} \varepsilon_{i1} \text{tr} \ G_{\tau_1}.
\]  
(6.13)

Here \(\varepsilon_{i1}\) is defined in (5.28). To ease the presentation, we further introduce the notation
\[
\tau_{i2} := -\tau_{i1} \text{tr} \: \tilde{B} G.
\]  
(6.14)

Using assumption (5.12), (5.19), and also (3.2), one checks that \(|\tau_{i1}| \ll 1, \ |\tau_{i2}| \ll 1, \text{ for all } i \in [1,N].

Similarly to (5.33), applying (5.32) to the first term on the right hand side of (6.12), we obtain
\[
\frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left[ \frac{1}{\|g_i\|} \frac{\partial (e_i^* \tilde{B}^{(i)} G e_i)}{\partial g_{ik}} \text{tr} \ G_{\tau_1} m^{(p-1,p)} \right] = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{k=1}^{N} \mathbb{E} \left[ \frac{1}{\|g_i\|} \frac{\partial (e_i^* \tilde{B}^{(i)} G e_i)}{\partial g_{ik}} \text{tr} \ G_{\tau_1} m^{(p-1,p)} \right]
\]
\[
+ \frac{1}{N^2} \sum_{i=1}^{N} \sum_{k=1}^{N} \mathbb{E} \left[ \frac{1}{\|g_i\|} e_i^* \tilde{B}^{(i)} G e_i \frac{\partial (\text{tr} \ G_{\tau_1})}{\partial g_{ik}} m^{(p-1,p)} \right]
\]
\[
+ \frac{1}{N^2} \sum_{i=1}^{N} \sum_{k=1}^{N} \mathbb{E} \left[ \frac{1}{\|g_i\|} e_i^* \tilde{B}^{(i)} G e_i \frac{\partial (\text{tr} \ G_{\tau_1})}{\partial g_{ik}} m^{(p-1,p)} \right]
\]
\[
+ \frac{1}{N^2} \sum_{i=1}^{N} \sum_{k=1}^{N} \mathbb{E} \left[ \frac{1}{\|g_i\|} e_i^* \tilde{B}^{(i)} G e_i \frac{\partial (\text{tr} \ G_{\tau_1})}{\partial g_{ik}} m^{(p-1,p)} \right].
\]  
(6.15)

First, we estimate the first term on the right hand side of (6.15). Using (5.50) and the bound
\[
\frac{1}{N} \sum_{i=1}^{N} \Pi_i^2 \leq 2\Pi^2,
\]
we have
\[
\frac{1}{N^2} \sum_{i=1}^{N} \sum_{k=1}^{N} \mathbb{E} \left[ \frac{1}{\|g_i\|} \frac{\partial (e_i^* \tilde{B}^{(i)} G e_i)}{\partial g_{ik}} \text{tr} \ G_{\tau_1} \right] = -\frac{1}{N} \sum_{i=1}^{N} \left( G_{i1} \text{tr} \tilde{B} G - (G_{i1} + T) \Upsilon \right) \tau_{i1}
\]
\[
+ \frac{1}{N^2} \sum_{i=1}^{N} \sum_{k=1}^{N} \left( \hat{T}_i - \frac{1}{\|g_i\|} \frac{\partial (e_i^* G e_i)}{\partial g_{ik}} \right) \tau_{i2} + \varepsilon_2 + O_{\prec}(\Pi^2),
\]
where we have introduced
\[
\varepsilon_2 := \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\|g_i\|} \tau_{i1} \varepsilon_{i2};
\]  
(6.16)

see (5.51) for the definition of \(\varepsilon_{i2}\). According to the definition in (6.11), we observe that
\[
\frac{1}{N} \sum_{i=1}^{N} \left( G_{i1} \text{tr} \tilde{B} G - (G_{i1} + T) \Upsilon \right) \tau_{i1} = \frac{1}{N^2} \sum_{i=1}^{N} G_{i1} \tau_{i1} (\text{tr} \tilde{B} G - \Upsilon) - \frac{1}{N} \sum_{i=1}^{N} T_i \tau_{i1} \Upsilon = O_{\prec}(\Psi \tilde{\Upsilon}).
\]

Here in the last step we used the facts
\[
\sum_{i=1}^{N} G_{i1} \tau_{i1} = 0, \quad \frac{1}{N} \sum_{i=1}^{N} T_i \tau_{i1} \Upsilon = O_{\prec}(\Psi \tilde{\Upsilon}),
\]  
(6.17)

where the second estimate is implied by the second estimate in (5.15), and the assumption that \(|\Upsilon| \ll \tilde{\Upsilon}.

Therefore, for the first term on the right hand side of (6.15), we have
\[
\frac{1}{N^2} \sum_{i=1}^{N} \sum_{k=1}^{N} \mathbb{E} \left[ \frac{1}{\|g_i\|} \frac{\partial (e_i^* \tilde{B}^{(i)} G e_i)}{\partial g_{ik}} \text{tr} \ G_{\tau_1} m^{(p-1,p)} \right]
\]
\[\begin{align*}
&= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{k} \mathbb{E} \left[ \left( \hat{T}_i \frac{1}{\| g_i \|} \frac{\partial (e_i^* G e_i)}{\partial g_{ik}} \right) \tau_{12} m^{(p-1,p)} \right] + \mathbb{E} \left[ (\varepsilon_2 + O_\prec(\Pi^2) + O_\prec(\Psi \Upsilon)) m^{(p-1,p)} \right] \\
&= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{k} \mathbb{E} \left[ \left( \frac{\partial \| g_i \|}{\partial g_{ik}} \right)^{-1} e_i^* G e_i \tau_{12} m^{(p-1,p)} \right] + \frac{1}{N^2} \sum_{i=1}^{N} \sum_{k} \mathbb{E} \left[ \frac{1}{\| g_i \|} \frac{\partial \tau_{12} e_i^* G e_i m^{(p-1,p)}}{\partial g_{ik}} \right] \\
&\quad + \frac{p-1}{N^2} \sum_{i=1}^{N} \sum_{k} \mathbb{E} \left[ \frac{1}{\| g_i \|} e_i^* G e_i \tau_{12} \left( \frac{1}{N} \sum_{j=1}^{N} \frac{\partial (d_j Q_j)}{\partial g_{ik}} \right) m^{(p-2,p)} \right] \\
&\quad + \frac{p}{N^2} \sum_{i=1}^{N} \sum_{k} \mathbb{E} \left[ \frac{1}{\| g_i \|} e_i^* G e_i \tau_{12} \left( \frac{1}{N} \sum_{j=1}^{N} \frac{\partial (d_j Q_j)}{\partial g_{ik}} \right) m^{(p-1,p-1)} \right] \\
&\quad + \mathbb{E} \left[ (\varepsilon_2 + O_\prec(\Pi^2) + O_\prec(\Psi \Upsilon)) m^{(p-1,p)} \right],
\end{align*}\]

where the second equation is obtained analogously to (5.54), by writing \( \hat{T}_i = \sum_k (g_{ik} e_i^* G e_i)/\| g_i \| \) and performing integration by parts with respect to the \( g_{ik} \)'s.

According to (6.12), (6.15), and (6.18), it suffices to estimate the last term on the right side of (6.12), the last four terms on the right side of (6.15), and all the terms on the right side of (6.18). All the desired estimates can be derived from the following lemma.

**Lemma 6.3.** Fix a \( z \in \mathcal{D}_r(\eta_n, \mu_m) \). 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 \( \hat{d}_1, \ldots, \hat{d}_N \in \mathbb{C} \) be any (possibly random) numbers with the bound \( \max_i |\hat{d}_i| < 1 \). Let \( Q \) be any (possibly random) diagonal matrix that satisfies \( \| Q \| < 1 \). Set \( X = I \) or \( A \), and set \( X_i = I \) or \( \hat{B}^{(i)} \). Then we have

\[\begin{align*}
&\frac{1}{N^2} \sum_{i=1}^{N} \sum_{k} \hat{d}_i \frac{\partial \| g_i \|}{\partial g_{ik}} e_i^* X_i G e_i = O_\prec \left( \frac{1}{N} \right), \\
&\frac{1}{N^2} \sum_{i=1}^{N} \sum_{k} \hat{d}_i \text{tr} \left( Q X_i \frac{\partial G}{\partial g_{ik}} \right) e_i^* X_i G e_i = O_\prec(\Psi^2 \Pi^2),
\end{align*}\]

and the same estimate holds if we replace \( \frac{\partial G}{\partial g_{ik}} \) by the complex conjugate \( \frac{\partial G^{\ast}}{\partial g_{ik}} \) in (6.20). Further, we have

\[\begin{align*}
\mathbb{E} \left[ \varepsilon_j m^{(p-1,p)} \right] &= \mathbb{E} \left[ O_\prec(\Pi^2) m^{(p-1,p)} \right] \\
&\quad + \mathbb{E} \left[ O_\prec(\Psi^2 \Pi^2) m^{(p-2,p)} \right] + \mathbb{E} \left[ O_\prec(\Psi^2 \Pi^2) m^{(p-1,p-1)} \right], \quad j = 1, 2.
\end{align*}\]

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

The second term of (6.15) and the first term of (6.18) are directly estimated by (6.19). Using the definition of \( \tau_{11} \) in (6.11) and of \( \tau_{12} \) in (6.14), the boundedness of the tracial quantities (c.f., (5.19)), and the chain rule, we get the estimate on the third term of (6.15) and the second term of (6.18), using (6.20) and the assumption (6.2). For the last two terms of (6.15), and the third and fourth terms of (6.18), we note that

\[\begin{align*}
\frac{1}{N} \sum_{j=1}^{N} d_j Q_j &= \text{tr} D \hat{B} G \text{tr} G - \text{tr} \hat{B} G \text{tr} DG = \text{tr} D \text{tr} G - \text{tr} DG - \text{tr} D A G \text{tr} G + \text{tr} A G \text{tr} DG,
\end{align*}\]

where in the last step we used the first identity of (5.8). Hence, by the chain rule, the fourth term of (6.15) and the third term of (6.18) are estimated with the aid of (6.20) and (6.2). The last term of (6.15) and the fourth term of (6.18) can be estimated analogously. Finally, the estimates of the second term of (6.12) and the last term of (6.18) are given by (6.21). Thus we conclude the proof of Lemma 6.2. \( \square \)

In the sequel, we prove Lemma 6.3.

**Proof of Lemma 6.3.** Note that (6.19) and (6.20) follow from the first and the last estimates in (5.55), respectively, by averaging over the index \( i \). Hence, it suffices to prove (6.21). Recall the definition of \( \varepsilon_1 \) from (6.13) and of \( \varepsilon_2 \) from (6.16).
We first consider $E[\varepsilon_1 \mathbf{m}^{(p-1,p)}]$. Recall the definition of $\varepsilon_1$ from (5.28). Using (5.14), (5.15), the first bound in (5.16), and (5.29), we have

$$\varepsilon_1 = \frac{\hat{h}_i^* \tilde{B}^i(h_i + O_\omega(\sqrt{\frac{1}{N}}) = \frac{\hat{h}_i^* \tilde{B}^i}{h_i + \omega_B^2} + O_\omega(\tilde{\Pi}^2).}$$  \hspace{1cm} (6.22)

Here the last step follows from the assumption $\frac{1}{\sqrt{N}} m < \tilde{\Pi}^2$, and that $h_i = \hat{h}_i + \frac{g_{\omega}}{\|\omega\|} e_i$, with

$$|\hat{g}_i| \sim \frac{1}{\sqrt{N}}, \quad \hat{h}_i^* \tilde{B}^i e_i = b_i \hat{h}_i e_i = 0.$$  

Hence, by the definition of $\varepsilon_1$ in (6.13), we have

$$\varepsilon_1 = \frac{1}{N} \sum_{i=1}^N \hat{h}_i^* \tilde{B}^i h_i \frac{d_i \text{tr} G - \text{tr} DG}{a_i - \omega_B^2} + O_\omega(\tilde{\Pi}^2) = \frac{1}{N} \sum_{i=1}^N \hat{h}_i^* \tilde{B}^i h_i \tau_{\tilde{\Pi}} + O_\omega(\tilde{\Pi}^2),$$

where we introduced the notation

$$\tau_{\tilde{\Pi}} := \frac{d_i \text{tr} G - \text{tr} DG}{a_i - \omega_B^2}.$$  

Using the integration by parts formula (5.32), we obtain

$$\frac{1}{N} \sum_{i=1}^N E[\hat{h}_i^* \tilde{B}^i h_i \tau_{\tilde{\Pi}} \mathbf{m}^{(p-1,p)}] = \frac{1}{N} \sum_{i=1}^N \sum_{k} \sum_{m} E \left[ \frac{1}{\|g_i\|^2} \hat{g}_i \bar{e}_k \bar{B}^i g_i \tau_{\tilde{\Pi}} \mathbf{m}^{(p-1,p)} \right]$$

$$\frac{1}{N^2} \sum_{i=1}^N \sum_{k} \sum_{m} \left[ \frac{\partial}{\partial g_{ik}} \left( \|g_i\|^{-2} \hat{e}_k \bar{B}^i g_i \tau_{\tilde{\Pi}} \mathbf{m}^{(p-1,p)} \right) \right].$$  \hspace{1cm} (6.23)

Note that

$$\frac{\partial}{\partial g_{ik}} \left( \|g_i\|^{-2} \hat{e}_k \bar{B}^i g_i \tau_{\tilde{\Pi}} \mathbf{m}^{(p-1,p)} \right) = \frac{\partial}{\partial g_{ik}} \left( \|g_i\|^{-2} \hat{e}_k \bar{B}^i g_i \tau_{\tilde{\Pi}} \mathbf{m}^{(p-1,p)} \right) + \|g_i\|^{-2} \hat{e}_k \bar{B}^i g_i \tau_{\tilde{\Pi}} \mathbf{m}^{(p-1,p)}$$

$$+ \|g_i\|^{-2} \hat{e}_k \bar{B}^i g_i \tau_{\tilde{\Pi}} \mathbf{m}^{(p-1,p)} + (p - 1)\|g_i\|^{-2} \hat{e}_k \bar{B}^i g_i \tau_{\tilde{\Pi}} \left( \sum_{j=1}^N \frac{\partial (d_j Q_j)}{\partial g_{ik}} \right) \mathbf{m}^{(p-2,p)}$$

$$+ p\|g_i\|^{-2} \hat{e}_k \bar{B}^i g_i \tau_{\tilde{\Pi}} \left( \sum_{j=1}^N \frac{\partial (d_j Q_j)}{\partial g_{ik}} \right) \mathbf{m}^{(p-1,p-1)}. \hspace{1cm} (6.24)$$

Notice that $\frac{\partial}{\partial g_{ik}} \left( \|g_i\|^{-2} \right) = -\|g_i\|^{-4} \bar{g}_{ik}$ and that $\tau_{\tilde{\Pi}} = O_\omega(1)$. In addition, we also have that

$$\sum_{k} \hat{g}_{ik} e_k = g_i^*, \quad \sum_{k} \hat{e}_k \bar{B}^i e_k = \text{Tr} B - b_i = b_i.$$  

Denoting by $\tilde{d}_1, \ldots, \tilde{d}_N \in \mathbb{C}$ generic (possibly random) numbers with $\max_i |\tilde{d}_i| < 1$, we see that the contributions from the first two terms on the right side of (6.24) to (6.23) follow from the estimates

$$\frac{1}{N^2} \sum_{i=1}^N \tilde{d}_i \hat{g}_i \bar{B}^i g_i = O_\omega \left( \frac{1}{N} \right), \quad \frac{1}{N^2} \sum_{i=1}^N \tilde{d}_i b_i \hat{e}_k \bar{B}^i e_k = O_\omega \left( \frac{1}{N} \right).$$

Here $\tilde{d}_i$ includes $\tau_{\tilde{\Pi}}$ and an appropriate power of $\|g_i\|$. In addition, for the estimate of the remaining terms in (6.24), we claim that, for $X_i = I, \bar{B}^i$,

$$\frac{1}{N^2} \sum_{i=1}^N \sum_{k} \tilde{d}_i \hat{e}_k X_i g_i \frac{\partial \tau_{\tilde{\Pi}}}{\partial g_{ik}} = O_\omega(\Psi^2 \Pi^2),$$  \hspace{1cm} (6.25)

$$\frac{1}{N^2} \sum_{i=1}^N \sum_{k} \tilde{d}_i \hat{e}_k X_i g_i \left( \frac{1}{N} \sum_{j=1}^N \frac{\partial (d_j Q_j)}{\partial g_{ik}} \right) = O_\omega(\Psi^2 \Pi^2),$$  \hspace{1cm} (6.26)

$$\frac{1}{N^2} \sum_{i=1}^N \sum_{k} \tilde{d}_i \hat{e}_k X_i g_i \left( \frac{1}{N} \sum_{j=1}^N \frac{\partial (d_j Q_j)}{\partial g_{ik}} \right) = O_\omega(\Psi^2 \Pi^2).$$  \hspace{1cm} (6.27)
The above three bounds follows from the last estimate in (5.55) and the chain rule. Hence, we conclude the proof of (6.21) with \( j = 1 \).

The proof of (6.21) for \( j = 2 \) is similar to \( j = 1 \). Recall the definition of \( \varepsilon_{12} \) from (5.51). Using (5.14), (5.15), the first bound in (5.16), and also the bounds in (5.29), we have

\[
\varepsilon_{12} = \left( \| g_i \|^2 - 1 \right) G_i \tr B G + O_{\sim} \left( \Psi \left( \frac{1}{N} \right) \right) = \left( \hat{g}_i^* \hat{g}_i - 1 \right) \frac{\tr B G}{a_i} - \omega_B^2 + O_{\sim} (\hat{\Pi}^2),
\]

which possesses a very similar structure as (6.22). The remaining proof is nearly the same as the case for \( \varepsilon_1 \); it suffices to replace \( \hat{g}_i^* B^{(i)} \hat{g}_i \) by \( \hat{g}_i^* \hat{g}_i \) throughout the proof. We thus omit the details. Hence, we conclude the proof for Lemma 6.3. \( \square \)

7. Optimal fluctuation averaging

In this section, we establish the optimal fluctuation averaging estimate for a very special linear combinations of the \( Q \)'s and their analogues the \( Q \)'s (c.f., (7.8)), under assumption (5.12).

Recall the definition of the approximate subordination functions \( \omega_\lambda^\ast \) and \( \omega_\lambda^\ast \) in (5.2). We denote

\[
\Lambda_A := \omega_\lambda^\ast - \omega_A, \quad \Lambda_B := \omega_B^\ast - \omega_B, \quad \Lambda := |\Lambda_A| + |\Lambda_B|. \tag{7.1}
\]

Recall \( S_{AB}, T_A \), and \( T_B \) defined in (3.1). For brevity, in the sequel, we use the shorthand notation \( S \equiv S_{AB} \).

**Proposition 7.1.** Fix a \( z = E + \eta \in D_\epsilon (\eta_\eta, \eta_\eta) \). Suppose that the assumptions of Proposition 5.1 hold. Suppose that \( \Lambda(z) \prec \hat{\Lambda}(z) \), for some deterministic and positive function \( \Lambda(z) \prec N^{-\frac{7}{8}} \), then

\[
\left| S \Lambda_i + T_{\Lambda_i^2} + O(\Lambda^2) \right| \leq \frac{\sqrt{\left( \Im \frac{m_{\mu_A \mu_B} + \hat{\Lambda}}{N\eta} \right) (|S| + \hat{\Lambda})}}{(N\eta)^2} + \frac{1}{(N\eta)^2}, \quad \tau = A, B. \tag{7.2}
\]

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.,

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

Indeed, observe that when \( \Lambda \prec \hat{\Lambda} \prec N^{-\frac{7}{8}} \), we obtain from the second line of (2.11) that

\[
|m_H - m_{\mu_A \mu_B}| = \left| m_H m_{\mu_A \mu_B} \right| \left| \frac{1}{m_H(z)} - \frac{1}{m_{\mu_A \mu_B}(z)} \right| \sim |m_H m_{\mu_A \mu_B}| \Lambda. \tag{7.4}
\]

Further, from the first line of (2.11) and (3.2), we see that, for any \( z \in D_\epsilon (\eta_\eta, \eta_\eta) \),

\[
|m_H m_{\mu_A \mu_B}| \prec \left| (m_{\mu_A \mu_B} + O_{\sim} (N^{-\frac{7}{8}})) m_{\mu_A \mu_B} \right| \prec 1. \tag{7.5}
\]

Hence, we conclude from (7.4) and (7.5) that

\[
|m_H - m_{\mu_A \mu_B}| \prec \Lambda \prec \hat{\Lambda}. \tag{7.6}
\]

Therefore, recalling (4.9), we have

\[
\Pi^2 \leq \frac{\Im \frac{m_{\mu_A \mu_B} + \hat{\Lambda}}{N\eta}}{N\eta} \prec \frac{\sqrt{\left( \Im \frac{m_{\mu_A \mu_B} + \hat{\Lambda}}{N\eta} \right) (|S| + \hat{\Lambda})}}{(N\eta)^2} \prec \Psi^2,
\]

where in the last two steps, we used that \( \Im m_{\mu_A \mu_B} / N \leq |S| < 1 \) (3.4) and (3.5). In addition, from (3.4) and (3.5), we also have \( \Im m_{\mu_A \mu_B} \leq |S| \geq \eta \). Thus we also have

\[
\frac{1}{N\sqrt{\eta}} \prec \frac{\sqrt{\left( \Im \frac{m_{\mu_A \mu_B} + \hat{\Lambda}}{N\eta} \right) (|S| + \hat{\Lambda})}}{N\eta}.
\]

From the definition of \( \Pi \) in (4.9), we note that up to a \( \frac{1}{\sqrt{\eta}} \) term \( \hat{\Pi} \) here is equivalent to \( \Pi \) inside the spectrum but it is much larger than \( \Pi \) in the outside regime where \( S \gg \Im m_{\mu_A \mu_B} \) (c.f., (3.4), (3.5)).

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

\[
\left| S \Lambda_i + T_{\Lambda_i^2} + O(\Lambda^2) \right| \prec \hat{\Pi}^2, \quad \tau = A, B. \tag{7.7}
\]
Recall the definition of $Q_i$ from (4.11). We also introduce their analogues
\[ Q_i \equiv Q_i(z) := (\tilde{A}G)_{ii} \text{tr} G - G_{ii} \text{tr} \tilde{A}G, \quad i \in [1, N]. \] (7.8)
with $\tilde{A}$ and $G$ given in (5.3). To prove Proposition 7.1, we need an optimal fluctuation averaging for a very special combination of $Q_i$’s and $Q_i$’s. To this end, we define the functions $\Phi_1, \Phi_2 : (C^+) \rightarrow C$,
\[ \Phi_1(\omega_1, \omega_2, z) := F_A(\omega_1) - \omega_1 - \omega_2 + z, \quad \Phi_2(\omega_1, \omega_2, z) := F_B(\omega_1) - \omega_1 - \omega_2 + z. \] (7.9)
From (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
\[ \Phi_1^i := \Phi_1(\omega_A^i, \omega_B^i, z), \quad \Phi_2^i := \Phi_2(\omega_A^i, \omega_B^i, z). \] (7.10)
Further, we define the quantities
\[ Z_1 := \Phi_1^i + (F_A^i(\omega_B) - 1) \Phi_2^i, \quad Z_2 := \Phi_2^i + (F_B^i(\omega_A) - 1) \Phi_1^i. \] (7.11)
We are going to show that $Z_1$ and $Z_2$ are actually certain linear combinations of the $Q_i$’s and the $Q_i$’s.
We start with the identities
\[ \Phi_1^i = - \frac{F_A(\omega_B)}{m_H(z)} \frac{1}{N} \sum_{i=1}^{N} a_i - \omega_B^i Q_i, \quad \Phi_2^i = - \frac{F_B(\omega_A)}{m_H(z)} \frac{1}{N} \sum_{i=1}^{N} b_i - \omega_A^i Q_i, \] (7.12)
which can be derived by combining (5.2), (5.4) and (5.58). For all $i \in [1, N]$, we set
\[ \eta_{i, 1} := - \frac{F_A(\omega_B)}{m_H(z)} \frac{1}{N} a_i - \omega_B^i, \quad \eta_{i, 2} := -(F_A(\omega_B) - 1) \frac{F_B(\omega_A)}{m_H(z)} \frac{1}{N} b_i - \omega_A^i. \] (7.13)
According to the definition in (7.11), (7.12), and also (7.13), we can write
\[ Z_1 = \frac{1}{N} \sum_{i=1}^{N} \eta_{i, 1} Q_i + \frac{1}{N} \sum_{i=1}^{N} \eta_{i, 2} Q_i, \] (7.14)
and $Z_2$ can be represented in a similar way.
Now, we choose $d_i = \eta_{i, 1}, i \in [1, N]$, in Proposition 6.1. Observe that $\eta_{i, 1}$ can be regarded as a smooth function of $\text{tr} \tilde{B}G = 1 - \text{tr} (A - z)G$ and $m_H(z) = \text{tr} G$, according to the definition in (7.13) and that of $\omega_B$ in (5.2). Then, using the chain rule and the estimates of the tracial quantities in (5.19), one can check that the first equation in assumption (6.2) is satisfied for the choice $d_i = \eta_{i, 1}, i \in [1, N]$, by using (5.55). The second equation can be checked analogously. Hence, applying Proposition 6.1, we get
\[ |\Phi_1| < \Psi \tilde{\Pi}, \quad |\Phi_2| < \Psi \tilde{\Pi}, \] (7.15)
where $\tilde{\Pi}$ is chosen as in (7.3).

The main technical task in this section is to establish the following estimates for $Z_1$ and $Z_2$, where the previous order $\Psi \tilde{\Pi}$ bounds from (6.3) are strengthened.

**Proposition 7.2.** Fix $z \in \mathcal{D}_\gamma(\eta_M, \eta_M)$. Suppose that the assumptions of Proposition 5.1 hold and that $\Lambda(z) \prec \tilde{\Lambda}(z)$ for some deterministic and positive function $\tilde{\Lambda}(z) \leq N^{-\frac{3}{2}}$. Choose $\tilde{\Pi}(z)$ as (7.3). Then,
\[ |Z_1| \prec \tilde{\Pi}^2, \quad |Z_2| \prec \tilde{\Pi}^2. \] (7.16)

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 $|\Lambda_A|, |\Lambda_B| \prec N^{-\frac{3}{2}}$. First of all, expanding $\Phi_1^i$ and $\Phi_2^i$ around $(\omega_A, \omega_B)$ and using the subordination equations $\Phi_1(\omega_A, \omega_B, z) = \Phi_2(\omega_A, \omega_B, z) = 0$, we get
\[ \Phi_1^i = -\Lambda_A + (F_A(\omega_B) - 1) \Lambda_B + \frac{1}{2} F_A''(\omega_B) \Lambda_B^2 + O(\Lambda_B^3), \]
\[ \Phi_2^i = -\Lambda_B + (F_B(\omega_A) - 1) \Lambda_A + \frac{1}{2} F_B''(\omega_A) \Lambda_A^2 + O(\Lambda_A^3). \] (7.17)
We rewrite the second equation in (7.17) as
\[ \Lambda_B = -\Phi_2^i + (F_B(\omega_A) - 1) \Lambda_A + \frac{1}{2} F_B''(\omega_A) \Lambda_A^2 + O(\Lambda_A^3). \] (7.18)
Substituting (7.18) into the first equation in (7.17) yields
\[ \Phi_1^i = -(F_A(\omega_B) - 1) \Phi_2^i + S_{22} \Lambda_A + T_{22} \Lambda_A^2 + O((\Phi_2^i)^2) + O(\Phi_2^i \Lambda_A) + O(\Lambda_A^3). \]
where $T_A$ is defined in (3.1). In light of the definition in (7.11), we have
\[
Z_1 = S \Lambda_A + T_A A_2^3 + O(\Phi_2^2) + O(\Phi_2^2 \Lambda_A) + O(\Lambda_A^3).
\]
Combination of (7.15), (7.16) with (7.19) leads to
\[
|S \Lambda_A + T_A A_2^3 + O(\Lambda_A^3)| < \hat{\Pi}^2 + \Psi \hat{\Pi} \Lambda.
\]
The second term on the right hand side of (7.20) can be absorbed into the first term, in light of the fact that $\Psi \Lambda \sim \hat{\Pi}$ (c.f., (7.3)). Hence, we have
\[
|S \Lambda_A + T_A A_2^3 + O(\Lambda_A^3)| < \hat{\Pi}^2.
\]
Analogously, we also have
\[
|S \Lambda_B + T_B A_2^3 + O(\Lambda_B^3)| < \hat{\Pi}^2.
\]
This completes the proof of Proposition 7.1.

It remains to prove Proposition 7.2. We state the proof for $Z_1$, $Z_2$ is handled similarly. We set
\[
t^{(k,l)} := \hat{Z}_k \hat{Z}_l^*, \quad k, l \in \mathbb{N}.
\]
We can now prove a stronger estimate one $E[t^{(p,p)}]$ 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 $z \in D_r(\eta_m, \eta_M)$. Suppose that the assumptions of Proposition 7.2 hold. For any fixed integer $p \geq 1$, we have
\[
E[t^{(p,p)}] = E[O_\prec(\hat{\Pi}^2)t^{(p-1,p)}] + E[O_\prec(\hat{\Pi}^4)t^{(p-2,p)}] + E[O_\prec(\hat{\Pi}^4)t^{(p-1,p-1)}].
\]
Now, with Lemma 7.3, we can prove Proposition 7.2.

**Proof of Proposition 7.2.** Similarly to the proof of (5.13) from Lemma 5.2, with Lemma 7.3, we can get (7.16) 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 $Z_1$ in (7.14). We can write
\[
E[t^{(p,p)}] = \frac{1}{N} \sum_{i=1}^{N} E[\mathcal{d}_{i,1} Q_i t^{(p-1,p)}] + \frac{1}{N} \sum_{i=1}^{N} E[\mathcal{d}_{i,2} Q_i t^{(p-1,p)}].
\]
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 (6.10), we can write
\[
\frac{1}{N} \sum_{i=1}^{N} \mathcal{d}_{i,1} Q_i = \frac{1}{N} \sum_{i=1}^{N} (\tilde{B} G)_{i,1} \text{tr} G \tau_{i,1},
\]
where we chose $d_i = \mathcal{d}_{i,1}, i \in [1, N]$, in the definition of $\tau_{i,1}$ in (6.11). Then, analogously to (6.12), we can also write
\[
\frac{1}{N} \sum_{i=1}^{N} E[\mathcal{d}_{i,1} Q_i t^{(p-1,p)}] = \frac{1}{N} \sum_{i=1}^{N} E[(\tilde{B} G)_{i,1} \text{tr} G \tau_{i,1} t^{(p-1,p)}],
\]
with $d_i = \mathcal{d}_{i,1}, i \in [1, N]$. Analogously to (6.5), we can show
\[
\frac{1}{N} \sum_{i=1}^{N} E[\mathcal{d}_{i,1} Q_i t^{(p-1,p)}] = E[O_\prec(\hat{\Pi}^2) t^{(p-1,p)}] + E[O_\prec(\Psi^2 \hat{\Pi}^2) t^{(p-2,p)}] + E[O_\prec(\Psi^2 \hat{\Pi}^2) t^{(p-1,p-1)}],
\]
where the last two terms come from the estimates of the analogues of the last two terms of (6.15), the third and fourth terms in the right side of (6.18), and also the terms in (6.26) and (6.27), but with $\frac{1}{N} \sum_{j=1}^{N} d_j Q_j$ replaced by $Z_1$. It suffices to improve the estimates of these terms. All these terms contain a derivative $\frac{\partial}{\partial g_{ik}}$ or $\frac{\partial}{\partial g_{ik}}$, which is smaller than the derivative of an arbitrary linear combination $\partial(\frac{1}{N} \sum_{j=1}^{N} d_j Q_j)/\partial g_{ik}$ or $\partial(\frac{1}{N} \sum_{j=1}^{N} d_j Q_j)/\partial g_{ik}$, due to the special choice of $\mathcal{d}_{i,1}$’s and $\mathcal{d}_{i,2}$’s. Specifically, we shall show the following lemma, which contains the estimates of all necessary terms.
Lemma 7.4. Fix a $z \in \mathcal{D}_r(\eta_{\text{in}}, \eta_{\text{M}})$. Suppose that the assumptions of Proposition 5.1 hold. Let $\bar{d}_1, \ldots, \bar{d}_N \in \mathbb{C}$ be (possibly random) numbers with $\max |\bar{d}_i| < 1$. Let $X_i = I$ or $\bar{B}^{(i)}$. Then we have
\[
\frac{1}{N^2} \sum_{i=1}^{N} \sum_{k} \bar{d}_i e_k^* X_i G e_k \frac{\partial G_{ik}}{\partial g_{ik}} = O_{\prec}(\hat{\Pi}^4),
\]
\[
\frac{1}{N^2} \sum_{i=1}^{N} \sum_{k} \bar{d}_i e_k^* X_i \bar{g} e_k \frac{\partial G_{ik}}{\partial g_{ik}} = O_{\prec}(\hat{\Pi}^4),
\]
\[
\frac{1}{N^2} \sum_{i=1}^{N} \sum_{k} \bar{d}_i e_k^* X_i \bar{g} e_k \frac{\partial G_{ik}}{\partial g_{ik}} = O_{\prec}(\hat{\Pi}^4).
\]
(7.24)

Proof of Lemma 7.4. We give the proof for the first estimate in (7.24). The third one is analogous, and the other two are just their complex conjugates. From the definitions in (7.10) and (7.11), we get
\[
\frac{\partial Z_1}{\partial g_{ik}} = \frac{\partial \Phi^i}{\partial g_{ik}} + (F_A^i(\omega_B) - 1) \frac{\partial \Phi^i}{\partial g_{ik}}
\]
\[
= \left((F_A^i(\omega_B) - 1)(F_B^i(\omega_A) - 1) - 1 \right) \frac{\partial \omega_A^i}{\partial g_{ik}} + (F_A^i(\omega_B) - 1) \frac{\partial \omega_B^i}{\partial g_{ik}}.
\]
Note that by the regularity of $F_A$ and $F_B$, we have
\[
(F_A^i(\omega_B) - 1)(F_B^i(\omega_A) - 1) - 1 = S + O(|A_A|), \quad F_A^i(\omega_B) - F_A^i(\omega_B) = O(|A_B|).
\]
The smallness of these coefficients carry the gain. According to the definition of $\hat{\Pi}$ in (7.3), we see that
\[
\langle |S| + \Lambda \Psi^2 \Pi^2 \leq \hat{\Pi}^4
\]
if $\Lambda \leq \tilde{\Lambda}$. Hence, for the first estimate in (7.24), it suffices to show that
\[
\frac{1}{N^2} \sum_{i=1}^{N} \sum_{k} \bar{d}_i e_k^* X_i G e_k \frac{\partial \omega^c}{\partial g_{ik}} = O_{\prec}(\Psi^2 \Pi^2), \quad \iota = A, B.
\]
(7.25)
This follows from (6.20), the fact that $\omega_B$ is a tracial quantity, and the chain rule. The other terms in (7.24) can be estimated similarly. This concludes the proof of Lemma 7.4. \qed

With the aid of Lemma 7.4, we can conclude the proof of Lemma 7.3. \qed

8. Strong local law

In this section, we use a continuity argument to prove the strong local law, i.e., Theorem 2.5, based on Propositions 5.1, 6.1, and 7.1. We start with the following lemma. Recall $S \equiv S_{AB}$ from (3.1) and $\Lambda = |A_A| + |A_B|$ from (7.1). Further recall that $\eta_m = N^{-\gamma}$, with $\gamma > 0$ as in Theorem 2.5.

Lemma 8.1. Fix $z \in \mathcal{D}_r(\eta_{\text{in}}, \eta_{\text{M}})$. Suppose that the assumptions of Proposition 5.1 hold. Let $\varepsilon \in (0, \frac{1}{12})$. Suppose that $\Lambda < \tilde{\Lambda}$ for some deterministic control parameter $\tilde{\Lambda} \leq N^{-\frac{7}{2}}$. If $\Lambda \geq \frac{N\varepsilon}{N\eta}$, then we have:
(i): If $\sqrt{\kappa + \eta} > N^{-\frac{\gamma}{2}}$, there is a sufficiently large constant $K_0 > 0$, such that
\[
1\left(\Lambda \leq \frac{|S|}{K_0}\right) |A_A| < N^{-2\varepsilon} \tilde{\Lambda}, \quad 1\left(\Lambda \leq \frac{|S|}{K_0}\right) |A_B| < N^{-2\varepsilon} \tilde{\Lambda};
\]
(8.1)
(ii): If $\sqrt{\kappa + \eta} \leq N^{-\frac{\gamma}{2}}$, we have
\[
|A_A| < N^{-\varepsilon} \tilde{\Lambda}, \quad |A_B| < N^{-\varepsilon} \tilde{\Lambda}.
\]
Proof. From (3.4) and (3.5), we see that $|S| \gtrsim \Im m_{\mu_A \mu_B}$ for all $z \in \mathcal{D}_r(\eta_{\text{in}}, \eta_{\text{M}})$. Thus (7.2) gives
\[
|S_{A_A} + T_A \Lambda^2 + O(\Lambda^3)| \prec \frac{|S| + \tilde{\Lambda}}{N\eta} + \frac{1}{(N\eta)^2}, \quad \iota = A, B.
\]
(8.2)
with $S$, $T_A$ and $T_B$ given in (3.1). Then, from $|A_A| \prec \Lambda \leq N^{-\frac{\gamma}{2}}$, we have
\[
S_{A_A} + T_A \Lambda^2 = O\left(\frac{|S| + \tilde{\Lambda}}{N\eta} + \frac{1}{(N\eta)^2} + N^{-\frac{\gamma}{2}} \tilde{\Lambda}^2\right), \quad \iota = A, B.
\]
(8.3)
If $\sqrt{\kappa + \eta} > N^{-\frac{\gamma}{2}} \tilde{\Lambda}$, we have for $\iota = A, B$,
\[
1\left(\Lambda \leq \frac{|S|}{K_0}\right) |A_A| \prec |S|^{-1}\left(\frac{|S| + \tilde{\Lambda}}{N\eta} + \frac{1}{(N\eta)^2} + N^{-\frac{\gamma}{2}} \tilde{\Lambda}^2\right) \leq C \frac{N}{N\eta} + N^{\varepsilon - \frac{\gamma}{2}} \tilde{\Lambda} \leq CN^{-2\varepsilon} \tilde{\Lambda}.
\]
(8.4)
Here we absorbed the quadratic term on the left hand side in (8.3) into the linear term. Hence, we proved (i). From (8.4), we also see that if $\sqrt{\kappa + \eta} > N^{-\varepsilon} \tilde{\Lambda}$, then

$$1 \left( \Lambda \leq \left| \frac{S}{K_0} \right| \right| \Delta_t \right| \propto N^{-\varepsilon}|S|, \quad t = A, B. \quad (8.5)$$

Next, we prove (ii). If $\sqrt{\kappa + \eta} \leq N^{-\varepsilon} \tilde{\Lambda}$, from (3.5) and (3.6), we see that $T_t \sim 1$. Hence, we solve the quadratic equation (8.3) directly, then we get

$$|\Delta_t| \propto |S| + \left( \frac{|S| + \tilde{\Lambda}}{N\eta} + \frac{1}{(N\eta)^2} + N^{-\varepsilon} \tilde{\Lambda}^2 \right)^{\frac{1}{2}} \leq CN^{-\varepsilon} \tilde{\Lambda}, \quad t = A, B,$$

under the assumption that $\tilde{\Lambda} \geq \frac{N^{2\varepsilon}}{N\eta}$. This concludes the proof of Lemma 8.1.

Recall the definitions of $S$ in (3.1) and of $A_d$, $\tilde{\Lambda}_d$, $\Lambda_T$, $\tilde{\Lambda}_T$ in (5.6). For any $z \in D_T(\eta_m, \eta_M)$ and any $\delta \in [0, 1]$, we define the event

$$\Theta(z, \delta) := \left\{ A_d(z) \leq \delta, \tilde{\Lambda}_d(z) \leq \delta, \Lambda(z) \leq \delta^2, \Lambda_T(z) \leq 1, \tilde{\Lambda}_T(z) \leq 1 \right\}. \quad (8.6)$$

We further decompose the domain $D_T(\eta_m, \eta_M)$ into the following two disjoint parts:

$$D_\geq := \left\{ z \in D_T(\eta_m, \eta_M) : \sqrt{\kappa + \eta} > \frac{N^{2\varepsilon}}{N\eta} \right\}, \quad D_\leq := \left\{ z \in D_T(\eta_m, \eta_M) : \sqrt{\kappa + \eta} \leq \frac{N^{2\varepsilon}}{N\eta} \right\}. \quad (8.7)$$

For $z \in D_\geq$, any $\delta \in [0, 1]$ and any $\varepsilon' \in [0, 1]$, we define the event $\Theta_\geq(z, \delta, \varepsilon') \subset \Theta(z, \delta)$ as

$$\Theta_\geq(z, \delta, \varepsilon') := \left\{ A_d(z) \leq \delta, \tilde{\Lambda}_d(z) \leq \delta, \Lambda(z) \leq \min\{\delta^2, N^{-\varepsilon}|S|\}, \Lambda_T(z) \leq 1, \tilde{\Lambda}_T(z) \leq 1 \right\}.$$

Lemma 8.2. Suppose that the assumptions in Theorem 2.5 hold. For any fixed $z \in D_T(\eta_m, \eta_M)$, any $\varepsilon \in (0, \frac{1}{12})$ and any $D > 0$, there exists a positive integer $N_1(D, \varepsilon)$ and an event $\Omega(z) \equiv \Omega(z, D, \varepsilon)$ with

$$P(\Omega(z)) \geq 1 - N^{-D}, \quad \forall N \geq N_1(D, \varepsilon) \quad (8.8)$$

such that the following hold:

(i) If $z \in D_\geq$, we have

$$\Theta_\geq \left( z, \frac{N^{2\varepsilon}}{\sqrt{N\eta}} \frac{\varepsilon}{10} \right) \cap \Omega(z) \subset \Theta_\geq \left( z, \frac{N^{2\varepsilon}}{\sqrt{N\eta}} \frac{\varepsilon}{2} \right). \quad (8.9)$$

(ii) If $z \in D_\leq$, we have

$$\Theta \left( z, \frac{N^{2\varepsilon}}{\sqrt{N\eta}} \right) \cap \Omega(z) \subset \Theta \left( z, \frac{N^{2\varepsilon}}{\sqrt{N\eta}} \right). \quad (8.10)$$

Proof. In this proof, we fix a $z \in D_T(\eta_m, \eta_M)$. From Proposition 5.1, we see that under the assumption

$$A_d(z) \sim N^{-\frac{\varepsilon}{2}}, \quad \tilde{\Lambda}_d(z) \sim N^{-\frac{\varepsilon}{2}}, \quad \Lambda_T(z) \sim 1, \quad \tilde{\Lambda}_T(z) \sim 1, \quad (8.11)$$

we have using (5.15) that

$$\Lambda(z) \sim \frac{1}{\sqrt{N\eta}}, \quad \tilde{\Lambda}_d(z) \sim \frac{1}{\sqrt{N\eta}}, \quad \Lambda_T(z) \sim \frac{1}{\sqrt{N\eta}}, \quad \tilde{\Lambda}_T(z) \sim \frac{1}{\sqrt{N\eta}}. \quad (8.12)$$

The following more quantitative statement for (8.12) can be derived if one states the proof of Proposition 5.1 in a quantitative way: if the event $\Theta(z, \frac{N^{2\varepsilon}}{\sqrt{N\eta}})$ holds, then

$$\Lambda(z) \leq \frac{N^{2\varepsilon}}{\sqrt{N\eta}}, \quad \tilde{\Lambda}_d(z) \leq \frac{N^{2\varepsilon}}{\sqrt{N\eta}}, \quad \Lambda_T(z) \leq \frac{N^{2\varepsilon}}{\sqrt{N\eta}}, \quad \tilde{\Lambda}_T(z) \leq \frac{N^{2\varepsilon}}{\sqrt{N\eta}}. \quad (8.13)$$

hold on $\Theta(z, \frac{N^{2\varepsilon}}{\sqrt{N\eta}}) \cap \Omega(z)$. Here $\Omega(z)$ is the typical “event” on which all the concentration estimates in the proof of Proposition 5.1 hold. Note that these concentration estimates are done with respect to the entries or quadratic forms of Gaussian vectors $g_j$’s, the probability of $\Omega(z)$ is thus independent of $z$. Hence, we have a positive integer $N_1(D, \varepsilon)$ uniformly in $z$ such that (8.8) holds. Moreover, on $\Omega(z)$, we can write Lemma 8.1 in a quantitative way. For instance, (8.1) can be written as $1 \left( \Lambda \leq \left| \frac{S}{K_0} \right| \right| \Delta_t \right| \leq N^{-\varepsilon} \tilde{\Lambda}$ on $\Omega(z)$.
Now, we choose \( \tilde{\Lambda} = \frac{N^{\epsilon}}{N\eta} \) in Lemma 8.1. From Lemma 8.1 (i) and (8.5), we see that for \( z \in \mathcal{D}_\prec \), the following bound holds on the event \( \Theta_\prec(z, \frac{N^{\epsilon}}{\sqrt{N\eta}}, \frac{N^{\epsilon}}{N\eta}) \cap \Omega(z) \),

\[
\Lambda \leq \min \left\{ \frac{N^{\epsilon}}{N\eta}, \frac{N^{-\epsilon} |S|}{N} \right\}. \tag{8.14}
\]

From Lemma 8.1 (ii), we see that for \( z \in \mathcal{D}_\preceq \), the following bound holds on the event \( \Theta(z, \frac{N^{\epsilon}}{\sqrt{N\eta}}, \frac{N^{\epsilon}}{N\eta}) \cap \Omega(z) \),

\[
\Lambda \leq \frac{N^{\epsilon}}{N\eta}. \tag{8.15}
\]

Substituting (8.14) and (8.15) into the first two estimates in (8.13), we further get that

\[
\Lambda_d(z) \leq \frac{N^{\epsilon}}{\sqrt{N\eta}}, \quad \tilde{\Lambda}_d(z) \leq \frac{N^{\epsilon}}{\sqrt{N\eta}}
\]

hold on \( \Theta_\prec(z, \frac{N^{\epsilon}}{\sqrt{N\eta}}, \frac{N^{\epsilon}}{N\eta}) \cap \Omega(z) \) if \( z \in \mathcal{D}_\prec \) and on \( \Theta(z, \frac{N^{\epsilon}}{\sqrt{N\eta}}, \frac{N^{\epsilon}}{N\eta}) \cap \Omega(z) \) if \( z \in \mathcal{D}_\preceq \). This completes the proof. \( \square \)

With Lemma 8.2, we can now prove (2.17) and (2.18) in Theorem 2.5, using a continuity argument. The proof of (2.19) will be stated in Section 9.

Proof of (2.17) and (2.18) in Theorem 2.5. With Lemma 8.2, the remaining proof of Theorem 2.5 is quite similar to the proof of Theorem 7.1 of [6]. So we only sketch the arguments.

We start with an entry-wise Green function subordination estimate on global scale, i.e., \( \eta = \eta_M \) for some sufficiently large constant \( \eta_M > 0 \). Recall \( Q_i \) from (4.11). We regard \( Q_i \) as a function of the random unitary matrix \( U \). Then, for \( z = E + i\eta_M \) with any fixed \( E \) and any \( \eta_M \geq \eta_M \), we apply the Gromov-Milman concentration inequality (c.f., (6.2) in [6]), and get

\[
|Q_i(E + i\eta_M) - \mathbb{E}Q_i(E + i\eta_M)| < \frac{1}{\sqrt{N\eta_M}}; \tag{8.16}
\]

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

\[
\mathbb{E}((\tilde{B}G \otimes G - G \otimes \tilde{B}G) = 0; \tag{8.17}
\]

see Proposition 3.2 of [24]. 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 (8.17) that

\[
\mathbb{E}Q_i = \mathbb{E}(\tilde{B}G)_{ii} \text{tr} G - \text{tr} \tilde{B}G = 0. \tag{8.18}
\]

We claim that, for sufficiently large \( \eta_M > 1 \), we have

\[
\sup_{z:1 \text{Im} z \geq \eta_M} |Q_i(z)| < \frac{1}{\sqrt{N}}, \quad \forall i \in [1,N], \tag{8.19}
\]

where we used (8.16), (8.18), the Lipschitz continuity of \( Q_i \) in the regime \( |z| \leq \sqrt{N} \) and the deterministic bound \( |Q_i(z)| \leq \frac{C}{N} \) when \( |z| \geq \sqrt{N} \). In addition, using that \( \|H\| \leq \|A\| + \|B\| < K \) and the convention \( \text{tr} \tilde{B} = \text{tr} B = 0 \) (c.f., (5.1)), we have, for \( z = E + i\eta_M \) with fixed \( E \) and any \( \eta_M \geq \eta_M \), the expansions

\[
\text{tr} G(z) = -\frac{1}{z} + O\left(\frac{1}{\eta_M^2}\right), \quad \text{tr} \tilde{B}G(z) = -\frac{\text{tr} \tilde{B}}{z} + O\left(\frac{1}{\eta_M^2}\right) = O\left(\frac{1}{\eta_M^2}\right), \tag{8.20}
\]

where we used \( \text{tr} B = 0 \) in the second equality. Hence, by the definition of \( \omega_{\tilde{B}} \) in (5.2), we see that

\[
\omega_{\tilde{B}}(z) = z + O\left(\frac{1}{\eta_M}\right), \quad z = E + i\eta_M. \tag{8.21}
\]

Using the identity \( (\tilde{B}G)_{ii} = 1 - (a_i - z)G_{ii} \), we can rewrite (8.19) as

\[
(1 - (a_i - \omega_{\tilde{B}})G_{ii}) \text{tr} G = O\left(\frac{1}{\sqrt{N}}\right), \quad z = E + i\eta_M. \tag{8.22}
\]

From the first line of (8.20) and (8.21) we get

\[
\Lambda_{\tilde{B}}^c(z) < \frac{1}{\sqrt{N}}, \quad z = E + i\eta_M. \tag{8.22}
\]
Analogously, we also have
\[ \tilde{\Lambda}_d(z) < \frac{1}{\sqrt{N}}, \quad z = E + i\eta_M. \] (8.23)

Averaging over the index \( i \) in the definition of \( \Lambda_i^c \) and \( \tilde{\Lambda}_i^c \) (c.f., (5.7)), using (8.22) and (8.23) and using the fact \( \text{tr} G = \text{tr} \tilde{G} = m_H \) yields
\[ \sup_{z : \text{Im} z \geq \eta_M} \left| m_H(z) - m_A(\omega_H(z)) \right| < \frac{1}{\sqrt{N}}, \quad \sup_{z : \text{Im} z \geq \eta_M} \left| m_H(z) - m_B(\omega_H(z)) \right| < \frac{1}{\sqrt{N}} \] (8.24)
where in the large \( z \) regime these bounds even hold deterministically, similarly to (8.19). This together with (5.4) gives us the system
\[ \sup_{z : \text{Im} z \geq \eta_M} \left| \Phi_1(\omega_A(z), \omega_H(z), z) \right| < \frac{1}{\sqrt{N}}, \quad \sup_{z : \text{Im} z \geq \eta_M} \left| \Phi_2(\omega_A(z), \omega_H(z), z) \right| < \frac{1}{\sqrt{N}}, \] (8.25)
where \( \Phi_1 \) and \( \Phi_2 \) are defined in (7.9). We regard (8.25) 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 \( \eta \) regime is analyzed in Lemma A.2. Choosing \( (\mu_1, \mu_2) = (\mu_A, \mu_B) \), \( \langle \hat{\omega}_1(z), \hat{\omega}_2(z) \rangle = \langle \omega_A(z), \omega_B(z) \rangle \) in Lemma A.2 below, and using the fact that (8.25) and (8.21) hold for any sufficiently large \( \eta_M \), we obtain from the stability Lemma A.2 that
\[ |\Lambda_i(z)| = |\omega_i^*(z) - \omega_i(z)| < \frac{1}{\sqrt{N}}, \quad i = A, B, \quad z = E + i\eta_M \] (8.26)
for any sufficiently large constant \( \eta_M > 1 \), say.

Substituting (8.26) into (8.22) and (8.23) gives
\[ \Lambda_d(E + i\eta_M) < \frac{1}{\sqrt{N}}, \quad \tilde{\Lambda}_d(E + i\eta_M) < \frac{1}{\sqrt{N}}, \] (8.27)
for any fixed \( E \in \mathbb{R} \). Using the bound \( \|G\| \leq \frac{1}{\sqrt{N}} \) and the inequality \( |x^* Gy| \leq \|G\| \|x\| \|y\| \), we also get
\[ \Lambda_T(E + i\eta_M) < \frac{1}{\eta_M}, \quad \tilde{\Lambda}_T(E + i\eta_M) < \frac{1}{\eta_M}, \] (8.28)
for any fixed \( E \in \mathbb{R} \). Since (8.27) and (8.28) guarantee assumption (5.12), similarly to (8.12), we can apply Proposition 5.1 to get, for any fixed \( E \in \mathbb{R} \), that
\[ \Lambda_T(E + i\eta_M) < \frac{1}{\sqrt{N}}, \quad \tilde{\Lambda}_T(E + i\eta_M) < \frac{1}{\sqrt{N}} \] (8.29)
Also observe that \( E + i\eta_M \in \mathcal{D}_\frac{3}{4} \), for any fixed \( E \), and that \( |S(E + i\eta_M)| \geq 1 \). Hence \( \Lambda(E + i\eta_M) \prec N^{-\epsilon}|S(E + i\eta_M)| \). Then we can apply Lemma 8.1 (i) repeatedly for smaller and smaller \( \Lambda \) to get
\[ \Lambda(E + i\eta_M) \prec \frac{1}{N}. \] (8.30)
Combining (8.27), (8.29), (8.30) with the fact \( \Lambda(E + i\eta_M) \prec N^{-\epsilon}|S(E + i\eta_M)| \), we see that the event \( \Theta_\>(E + i\eta_M, \frac{N^{\frac{3}{4}+\epsilon}}{N^\eta}, \frac{\epsilon}{10}) \) holds with high probability. More quantitatively, we have for any fixed \( E \) that
\[ \mathbb{P}\left(\Theta_\>(E + i\eta_M, \frac{N^{\frac{3}{4}+\epsilon}}{N^\eta}, \frac{\epsilon}{10}) \right) \geq 1 - N^{-D}, \] (8.31)
for all \( D > 0 \) and \( N \geq N_2(D, \epsilon) \) with some threshold \( N_2(D, \epsilon) \).

Now we take (8.31) as the initial input, and use a continuity argument based on Lemma 8.2, to control the probability of the “good” events \( \Theta_\> \) for \( z \in \mathcal{D}_\> \), and \( \Theta \) for \( z \in \mathcal{D}_\< \). To this end, we first recall the event \( \Omega(z) \) in Lemma 8.2. The main task is to show for any \( z = E + i\eta \in \mathcal{D}_\> \),
\[ \Theta_\>(E + i\eta, \frac{N^{\frac{3}{4}+\epsilon}}{N^\eta}, \frac{\epsilon}{2}) \cap \Omega(E + i(\eta - N^{-5})) \subset \Theta_\>(E + i(\eta - N^{-5}), \frac{N^{\frac{3}{4}+\epsilon}}{N^\eta}, \frac{\epsilon}{2}), \] (8.32)
and, for any \( z = E + i\eta \in \mathcal{D}_\< \),
\[ \Theta(E + i\eta, \frac{N^{\frac{3}{4}+\epsilon}}{N^\eta}) \cap \Omega(E + i(\eta - N^{-5})) \subset \Theta(E + i(\eta - N^{-5}), \frac{N^{\frac{3}{4}+\epsilon}}{N^\eta}). \] (8.33)
The inclusions (8.32) and (8.33) are analogous to (7.20) of [4]. The only difference is here we decompose the domain \( \mathcal{D}_\>(\eta_M, \eta_M) \) into \( \mathcal{D}_\> \) and \( \mathcal{D}_\< \), and in \( \mathcal{D}_\> \) we also keep monitoring the event \( \Lambda \leq N^{-\frac{3}{4}}|S| \) in
order to use Lemma 8.1 (i). As we are gradually reducing \( \text{Im} \ z \), once \( z \) enters into the domain \( \mathcal{D}_\leq \), we do not need to monitor \( S \) anymore.

The proofs of (8.32) and (8.33) rely on the Lipschitz continuity of the Green function, \( \|G(z) - G(z')\| \leq N^2|z - z'| \), and of the subordination functions and \( S \) in (3.7). Using the Lipschitz continuity of these functions, it is not difficult to see the following two

\[
\Theta_\geq \left( E + i\eta, \frac{N^{3/2} \varepsilon}{\sqrt{N\eta}} \frac{\varepsilon}{2} \right) \subset \Theta_\geq \left( E + i(\eta - N^{-5}), \frac{N^{3/2} \varepsilon}{\sqrt{N\eta}} \frac{\varepsilon}{10} \right), \quad z = E + i\eta \in \mathcal{D}_{\geq},
\]

(8.34)

\[
\Theta \left( E + i\eta, \frac{N^{3/2} \varepsilon}{\sqrt{N\eta}} \right) \subset \Theta \left( E + i(\eta - N^{-5}), \frac{N^{3/2} \varepsilon}{\sqrt{N\eta}} \right),
\]

(8.35)

Then, (8.34) together with (8.9) implies (8.32). Similarly, (8.35) together with (8.10) implies (8.33). Applying (8.32) and (8.33) recursively and using the simple fact that the domains \( \mathcal{D}_{\geq} \) and \( \mathcal{D}_\leq \) are connected, one can go from \( \eta = \eta_m \) to \( \eta = \eta_m \), step by step of size \( N^{-5} \). Consequently, we obtain for any \( \eta \in [\eta_m, \eta_M] \cap N^{-5}\mathbb{Z} \) that, if \( E + i\eta \in \mathcal{D}_{\geq} \) then

\[
\Theta_\geq \left( E + i\eta_m, \frac{N^{3/2} \varepsilon}{\sqrt{N\eta_m}} \frac{\varepsilon}{2} \right) \cap \Omega(E + i(\eta_M - N^{-5})) \cap \ldots \cap \Omega(E + i\eta) \subset \Theta_\geq \left( E + i\eta, \frac{N^{3/2} \varepsilon}{\sqrt{N\eta}} \right),
\]

(8.36)

respectively, if \( E + i\eta \in \mathcal{D}_\leq \) then

\[
\Theta_\geq \left( E + i\eta_m, \frac{N^{3/2} \varepsilon}{\sqrt{N\eta_m}} \right) \cap \Omega(E + i(\eta_M - N^{-5})) \cap \ldots \cap \Omega(E + i\eta) \subset \Theta \left( E + i\eta, \frac{N^{3/2} \varepsilon}{\sqrt{N\eta}} \right),
\]

(8.37)

Combining (8.8), (8.31), (8.36) and (8.37), we have

\[
\mathbb{P} \left( \Theta \left( E + i\eta, \frac{N^{3/2} \varepsilon}{\sqrt{N\eta}} \right) \right) \geq 1 - N^{-D}(1 + N^5(\eta_M - \eta)),
\]

(8.38)

uniformly for all \( \eta \in [\eta_m, \eta_M] \cap N^{-5}\mathbb{Z} \), when \( N \geq \max \{N_1(D, \varepsilon), N_2(D, \varepsilon)\} \). 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 \( \mathcal{D}_r(\eta_m, \eta_M) \).

By the definition in (8.6), we obtain from (8.38) that

\[
\begin{align*}
\max_{i \in [1, N]} \left| G_{ii}(z) - \frac{1}{a_i - \omega_B(z)} \right| &< \frac{1}{\sqrt{N\eta}}, & |\Lambda_A(z)| &< \frac{1}{N\eta}, \\
\max_{i \in [1, N]} G_{ii}(z) &> \frac{1}{b_i - \omega_A(z)} \frac{1}{\sqrt{N\eta}}, & |\Lambda_B(z)| &< \frac{1}{N\eta},
\end{align*}
\]

(8.39)

uniformly on \( \mathcal{D}_r(\eta_m, \eta_M) \) with high probability. For any deterministic \( d_1, \ldots, d_N \in \mathbb{C} \), we further write

\[
\frac{1}{N} \sum_{i=1}^N d_i \left( G_{ii} - \frac{1}{a_i - \omega_B^\prime} \right) = \frac{1}{N} \sum_{i=1}^N \frac{d_i}{\text{tr} \ G(a_i - \omega_B^\prime)} \bar{Q}_i,
\]

(8.40)

which can easily be checked from the definition of \( \omega_B^\prime \), \( Q_i \) and the equation \((a_i - z) G_{ii} + \tilde{B} G_{ii} = 1\). Regarding \( \frac{d_i}{\text{tr} \ G(a_i - \omega_B^\prime)} \) 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 (5.55). Hence, we have

\[
\left| \frac{1}{N} \sum_{i=1}^N d_i \left( G_{ii} - \frac{1}{a_i - \omega_B^\prime} \right) \right| < \Psi \bar{\Pi}.
\]

(8.41)

Plugging the last estimate in (8.39) into (8.41), and using (3.2), we obtain (2.17) uniformly on \( \mathcal{D}_r(\eta_m, \eta_M) \). Finally, choosing \( d_i = 1 \) for all \( i \in [1, N] \) in (8.41), we get (2.18) uniformly on \( \mathcal{D}_r(\eta_m, \eta_M) \). This completes the proof of (2.17) and (2.18) in Theorem 2.5. \qed
9. Rigidity of the eigenvalues

In this section, we prove Theorem 2.6, and also (2.19) in Theorem 2.5. Recall the definition of $D_\succ$ in (8.7). We start by improving the estimate of $\Lambda$ defined in (7.1) in the following subdomain of $D_\succ$,

$$D_\succ := \{ z = E + i\eta \in D_\succ : E < E_\succ \}.$$  \hfill (9.1)

**Lemma 9.1.** Suppose that the assumptions in Theorem 2.5 hold. Then, we have the following uniform estimate for all $z \in \tilde{D}_\succ$,

$$\Lambda(z) \leq \frac{1}{N\sqrt{(\kappa + \eta)\eta}} + \frac{1}{\sqrt{\kappa + \eta}(N\eta)^2}.$$  \hfill (9.2)

**Proof.** First, from (8.39), we see that $\Lambda \prec \frac{1}{N\eta}$ on $D_\tau(\eta_n, \eta_m)$. Now, suppose that $\Lambda \prec \Lambda$ for some deterministic $\Lambda \equiv \Lambda(z)$ that satisfies

$$N^\varepsilon \left( \frac{1}{N\sqrt{(\kappa + \eta)\eta}} + \frac{1}{\sqrt{\kappa + \eta}(N\eta)^2} \right) \leq \Lambda(z) \leq N^\varepsilon \frac{1}{N\eta}. \hfill (9.3)$$

Observe that such $\Lambda$ always exists on $D_\succ$. From (7.2), (3.4) and (3.5), we have for $\iota = A, B$, and $z \in \tilde{D}_\succ$

$$|S\Lambda_\iota + \tau_\iota\Lambda_\iota^2| \prec \sqrt{\frac{\eta}{\sqrt{\kappa + \eta} + \Lambda}(\sqrt{\kappa + \eta} + \Lambda)} \frac{1}{N\eta} + \frac{1}{(N\eta)^2} \prec \sqrt{\Lambda\sqrt{\kappa + \eta}} \frac{\sqrt{\eta}}{N\eta} + \frac{1}{(N\eta)^2}, \hfill (9.4)$$

where we used that $\Lambda \prec \frac{N^\varepsilon}{N\eta} \leq N^{-\varepsilon}\sqrt{\kappa + \eta}$ for all $z \in \tilde{D}_\succ$. Moreover, for $z \in \tilde{D}_\succ$, we see that

$$|\Lambda_\iota| \prec \frac{1}{N\eta} \leq N^{-2\varepsilon}\sqrt{\kappa + \eta} \sim N^{-2\varepsilon}|S|,$$

for $\iota = A, B$. Hence, according to the fact $\tau_\iota \leq C$ (c.f., (3.5)), we can absorb the second term on the left side of (9.4) into the first term, and thus we have for $\iota = A, B$

$$|\Lambda_\iota| \prec \frac{1}{\sqrt{\kappa + \eta}} \left( \frac{\sqrt{\Lambda}\sqrt{\kappa + \eta} \frac{\sqrt{\eta}}{N\eta} + \frac{1}{(N\eta)^2}}{N\eta} \right) \leq \frac{1}{N\eta(\kappa + \eta)^{\frac{1}{2}}} \Lambda \prec \Lambda \leq N^{-\varepsilon} \Lambda,$$

where in the second step we used the lower bound in (9.3) directly, and in the last step we used the fact $(N\eta)^{-1}(\kappa + \eta)^{-\frac{1}{2}} \leq N^{-\varepsilon}\Lambda^\frac{1}{2}$ which again follows from the lower bound in (9.3).

Hence, we improved the bound from $\Lambda \leq \Lambda$ to $\Lambda \leq N^{-\varepsilon}\Lambda$ as long as the lower bound in (9.3) holds. Performing the above improvement iteratively, one finally gets (9.2). Hence, we complete the proof. $\square$

With the aid of Lemma 9.1, we can now prove Theorem 2.6.

**Proof of Theorem 2.6.** We first show (2.21) for the smallest eigenvalue $\lambda_1$, i.e.,

$$|\lambda_1 - \gamma_1| \prec N^{-\varepsilon} \frac{1}{\gamma_1}.$$  \hfill (9.5)

Recall $K$ defined in (2.13). For any (small) constant $\varepsilon > 0$, we define the line segment.

$$\tilde{D}(\varepsilon) := \{ z = E + i\eta : E \in [-K, E_\succ - N^{-\varepsilon}\frac{\gamma_1}{\varepsilon}], \eta = N^{-\varepsilon}\frac{\gamma_1}{\varepsilon} \}. \hfill (9.6)$$

Then it is easy to check that $\tilde{D}(\varepsilon) \subset \tilde{D}_\succ$ (c.f., (9.1)). Applying (9.2), we obtain $\Lambda \prec \frac{N^{-\varepsilon}}{N\eta}$ uniformly on $\tilde{D}(\varepsilon)$, which together with (7.6) implies

$$|m_H(z) - m_{\mu, \Lambda}(z)| \prec \frac{N^{-\varepsilon}}{N\eta}, \hfill (9.7)$$

uniformly on $\tilde{D}(\varepsilon)$. Moreover, by (3.4), we have

$$\text{Im} m_{\mu, \Lambda}(z) \sim \frac{\eta}{\sqrt{\kappa + \eta}} \leq \frac{N^{-\varepsilon}}{N\eta}, \hfill (9.8)$$

uniformly on $\tilde{D}(\varepsilon)$. Combining (9.7) with (9.8) yields

$$\text{Im} m_H(z) \prec \frac{N^{-\varepsilon}}{N\eta}, \hfill (9.9)$$
uniformly on \( \overline{D}(\varepsilon) \). Since \( \|H\| < K \), to see (9.5), it suffices to show that with high probability \( \lambda_1 \) is not in the interval \([−K, E_- − N^{−\frac{5}{2} + \varepsilon}]\). We prove it by contradiction. Suppose that \( \lambda_1 \in [−K, E_- − N^{−\frac{5}{2} + \varepsilon}] \). Then clearly for any \( \eta > 0 \),
\[
\sup_{E \in [−K, E_- − N^{−\frac{5}{2} + \varepsilon}]} \text{Im } m_H(E + i\eta) = \sup_{E \in [−K, E_- − N^{−\frac{5}{2} + \varepsilon}]} \frac{1}{N} \sum_{i=1}^{N} \frac{\eta}{(\lambda_i - E)^2 + \eta^2} \geq \frac{1}{N\eta},
\]
which contradicts the fact that (9.9) holds uniformly on \( \overline{D}(\varepsilon) \). Hence, we have (9.5).

Next, from (2.18), (3.80) and (3.81) and a standard application of Helffer-Sjöstrand formula (c.f., Lemma 5.1 [2]) on \( D_\tau(\eta_m, \eta_M) \) yields
\[
\sup_{z \leq E_- + c} |\mu_H((−\infty, x]) - \mu_A \Box \mu_B((−\infty, x])| \lesssim \frac{1}{N},
\]
for any sufficiently small \( c = c(\tau) \). Then (9.5), (9.10), together with the rigidity (3.94) and the square root behavior of the distribution \( \mu_A \Box \mu_B \) (c.f., (3.62)) will lead to the conclusion. The same conclusion holds with \( \gamma_j \)'s replaced by \( \gamma_j \)'s by rigidity (3.94).

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 \( \kappa = E_- − E \geq N^{−\frac{3}{2} + \varepsilon} \) in (2.19), we see that (2.19) follows from (2.18) directly in the regime \( \eta \geq \frac{k}{4} \), say. Hence, in the sequel, we work in the regime \( \eta \leq \frac{k}{4} \) only. For any \( z = E + i\eta \in D_\tau(\eta_m, \eta_M) \) with \( \kappa \geq N^{−\frac{3}{2} + \varepsilon} \), we set the contour \( C \equiv C(z) := C_1 \cup C_2 \cup C_3 \cup \overline{C_3} \),
where
\[
C_1 \equiv C_1(z) := \{ \tilde{z} = E + \frac{K}{2} + i\eta : -\eta - \kappa \leq \eta \leq \eta + \kappa \}, \quad C_2 \equiv C_2(z) := \{ \tilde{z} = E - \frac{K}{2} + i\eta : -\eta - \kappa \leq \eta \leq \eta + \kappa \}, \quad C_3 \equiv C_3(z) := \{ \tilde{z} = E \pm i(\eta + \kappa) : E - \frac{K}{2} \leq E \leq E + \frac{K}{2} \}
\]
We then further decompose \( C = C_< \cup C_\geq \), where \( C_< \equiv C_< \equiv \{ \tilde{z} \in C : |\text{Im } \tilde{z}| < \eta_m \} \), \( C_\geq \equiv C_\geq \equiv \{ \tilde{z} \in C : |\text{Im } \tilde{z}| > \eta_m \} \).

Now, we further introduce the event
\[
\Xi := \bigcap_{\tilde{z} \in C_<} \{ |m_H(\tilde{z}) - m_{\mu_A \Box \mu_B}(\tilde{z})| \leq \frac{N^\varepsilon}{N\text{Im } \tilde{z}} \} \cap \{ \lambda_1 \geq E_1 = \frac{1}{4}N^{−2/3 + \varepsilon} \}
\]
Then, on the event \( \Xi \), we have
\[
m_H(z) - m_{\mu_A \Box \mu_B}(z) = \frac{1}{2\pi i} \int_C \frac{1}{\tilde{z} - z} (m_H(\tilde{z}) - m_{\mu_A \Box \mu_B}(\tilde{z})) d\tilde{z} = \frac{1}{2\pi i} \left( \int_{C_<} + \int_{C_> \kappa} + \int_{C_> \varepsilon} \right) \frac{1}{\tilde{z} - z} (m_H(\tilde{z}) - m_{\mu_A \Box \mu_B}(\tilde{z})) d\tilde{z}.
\]
Note that, for \( \tilde{z} \in C_< \), we always have \( \frac{1}{|\tilde{z} - z|} \leq \frac{2}{\kappa} \). In addition, for \( \tilde{z} \in C_< \), we have the fact \( |C_<| \leq \eta_m \), and
\[
|m_H(\tilde{z})| \leq \frac{C}{\kappa}, \quad |m_{\mu_A \Box \mu_B}(\tilde{z})| \leq \frac{C}{\kappa},
\]
which hold on \( \Xi \). For \( \tilde{z} \in C_\geq \), we have the fact \( |C_\geq| \leq C_\kappa \) and the bound
\[
|m_H(\tilde{z}) - m_{\mu_A \Box \mu_B}(\tilde{z})| \leq \frac{N^\varepsilon}{N\text{Im } \tilde{z}}
\]
which holds on \( \Xi \). Applying the above bounds to (9.11), it is elementary to check that
\[
|m_H(z) - m_{\mu_A \Box \mu_B}(z)| \leq C(\eta_m + N^{−1 + \varepsilon} \log N) \frac{1}{\kappa}
\]
on \( \Xi \). Since \( \gamma \in \eta_m = N^{−1+\gamma} \) and \( \varepsilon \) can be arbitrary, we can conclude that
\[
|m_H(z) - m_{\mu_A \Box \mu_B}(z)| < \frac{1}{N\kappa}
\]
(9.12)
if we can show that $\Xi$ holds with high probability. Using (9.5), it suffices to show that
\[ |m_H(\tilde{z}) - m_{\mu, \Xi}(\tilde{z})| < \frac{1}{N \text{Im} \tilde{z}}, \]
uniformly in $\tilde{z} \in C_\ast$. This only requires us to enlarge the domain $D_r(\eta_m, \eta_M)$ and also consider its complex conjugate to include $C_\ast$ during the proof of (2.18). Hence, we conclude the proof of (2.19) by combining the $\frac{1}{N \eta}$ bound in (9.12) 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 $\sup_{\mathbb{C}^+} |m_{\mu, \Xi}| \leq C$ (c.f., (3.8)), 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, c.f., Theorem 2.4 in [5], we have
\[ |m_H(z) - m_{\mu, \Xi}(z)| < \frac{1}{N \eta}, \quad (9.13) \]
uniformly on the domain $D(\eta_m, \eta_M) := \{ z = E + i\eta \in \mathbb{C}^+ : -K \leq E \leq K, \quad \eta_m \leq \eta \leq \eta_M \}$. Then, (9.13) 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.

\[ \square \]

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 $y = (y_i) \in \mathbb{C}^N$ be a deterministic complex vector. For a Gaussian random vector $g = (g_1, \ldots, g_N) \in \mathcal{N}_G(0, \sigma^2 I_N)$ or $\mathcal{N}_C(0, \sigma^2 I_N)$, we have
\[ |y^* g| \prec \sigma|y|, \quad |g^* X g - \sigma^2 N \text{tr} X| \prec \sigma^2 \|X\|_2. \quad (A.1) \]

A.2. Stability for large $\eta$. For any probability measures $\mu_1$ and $\mu_2$ on the real line, we define the functions $\Phi_1, \Phi_2 : (\mathbb{C}^+) \to \mathbb{C}$ by setting
\[ \Phi_1(\omega_1, \omega_2, z) := F_{\mu_1}(\omega_2) - \omega_1 - \omega_2 + z, \quad \Phi_2(\omega_1, \omega_2, z) := F_{\mu_2}(\omega_1) - \omega_1 - \omega_2 + z. \quad (A.2) \]

We observe that the system of subordination equations (2.9) is equivalent to
\[ \Phi_1(\omega_1(z), \omega_2(z), z) = 0, \quad \Phi_1(\omega_1(z), \omega_2(z), z) = 0, \quad \forall z \in \mathbb{C}^+. \]

We have the following linear stability for the subordination equation in the large $\eta$ 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 $\tilde{\eta}_0 > 0$ be any (large) positive number and let $\tilde{\omega}_1, \tilde{\omega}_2, \tilde{r}_1, \tilde{r}_2 : C_{\tilde{\eta}_0} \to \mathbb{C}$ be analytic functions where $C_{\tilde{\eta}_0} := \{ z \in \mathbb{C} : \text{Im} z \geq \tilde{\eta}_0 \}$. Assume that there is a constant $C > 0$ such that the following hold for all $z \in C_{\tilde{\eta}_0}$:
\[ |\text{Im} \tilde{\omega}_1(z) - \text{Im} z| \leq C, \quad |\text{Im} \tilde{\omega}_2(z) - \text{Im} z| \leq C, \quad (A.3) \]
\[ |\tilde{r}_1(z)| \leq C, \quad |\tilde{r}_2(z)| \leq C, \quad (A.4) \]
\[ \Phi_1(\tilde{\omega}_1(z), \tilde{\omega}_2(z), z) = \tilde{r}_1(z), \quad \Phi_2(\tilde{\omega}_1(z), \tilde{\omega}_2(z), z) = \tilde{r}_2(z). \quad (A.5) \]
Then there is a constant $\eta_0$ with $\eta_0 \geq \bar{\eta}_0$, such that
\[
|\tilde{\omega}_1(z) - \omega_1(z)| \leq 2\|\tilde{r}(z)\|,
\]
\[
|\tilde{\omega}_2(z) - \omega_2(z)| \leq 2\|\tilde{r}(z)\|,
\]
on the domain $\mathbb{C}_{\eta_0} := \{z \in \mathbb{C} : \text{Im } z \geq \eta_0\}$, where $\omega_1(z)$ and $\omega_2(z)$ are the subordination functions associated with $\mu_1$ and $\mu_2$.\hspace{1cm} (A.6)

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 $\tilde{\omega}_1(z), \tilde{\omega}_2(z)$ such that $\Phi_{\mu_1,\mu_2}(\tilde{\omega}_1(z), \tilde{\omega}_2(z), z) = 0$ with
\[
|\tilde{\omega}_j(z) - \bar{\omega}_j(z)| \leq 2\|\tilde{r}(z)\|,
\]
\[
j = 1, 2, \quad z \in \mathbb{C}_{\eta_0}.
\]
From (4.20) of [3] we know that the Jacobian of the subordination equations (denoted by $\Gamma_{\mu_1,\mu_2}$ in [3]) is close to 1 for sufficiently large $\bar{\eta}_0$. Thus by analytic inverse function theorem we obtain that $\tilde{\omega}_j(z), j = 1, 2$, are also analytic functions for large $\eta = \text{Im } z$. From (A.3), (A.4) and (A.7), we see that
\[
\lim_{\eta \nearrow \infty} \frac{\text{Im } \tilde{\omega}_1(\eta \cdot i)}{\eta} = \lim_{\eta \nearrow \infty} \frac{\text{Im } \tilde{\omega}_2(\eta \cdot i)}{\eta} = 1.
\]
It is known from the proof of the uniqueness of the solution to the subordination equations near $z = i\infty$ that $(\tilde{\omega}_1(z), \tilde{\omega}_2(z))$ is the unique solution in a neighborhood of $z = i\infty$ and it can be analytically extended to all $z \in \mathbb{C}^+$. Hence, $(\tilde{\omega}_1(z), \tilde{\omega}_2(z)) = (\omega_1(z), \omega_2(z))$. This together with (A.7) concludes the proof. \hspace{1cm} \square

Appendix B.

In this appendix, we prove some technical lemmas. First, we estimate the small terms involving $\Delta_G$. Specifically, we provide the bounds for the $\Delta_G$ involved terms in the last four estimates in Lemma 5.3. Then, we prove Lemma 5.3. We summarize the estimates for $\Delta_G$ involved terms in the following lemma.

Lemma B.1. Fix a $z \in \mathcal{D}_T(\eta_0, \eta_M)$. Let $Q \in M_N(\mathbb{C})$ be arbitrary, with $\|Q\| < 1$. Let $X = I$ or $\tilde{B}^{(i)}$, and $X = I$ or $A$. Suppose the assumptions of Proposition 5.1 hold. Then, we have
\[
\frac{1}{N} \sum_{k}^{(i)} e_{\phi}^{*} X \Delta_G(i, k) e_{\phi} = O_{\epsilon}(\Pi_{\epsilon}),
\]
\[
\frac{1}{N} \sum_{k}^{(i)} e_{\phi}^{*} X \Delta_G(i, k) e_{\phi} X G e_{\phi} = O_{\epsilon}(\Pi_{\epsilon}),
\]
\[
\frac{1}{N} \sum_{k}^{(i)} h_{\psi}^{*} X \Delta_G(i, k) e_{\phi} X G e_{\phi} = O_{\epsilon}(\Pi_{\epsilon}),
\]
\[
\frac{1}{N} \sum_{k}^{(i)} \text{tr } Q X \Delta_G(i, k) e_{\phi} X G e_{\phi} = O_{\epsilon}(\Psi^{2} \Pi_{\epsilon}).
\]
\hspace{1cm} (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 (5.39) and (5.38). We note that $\Delta_R(i, k)$ is a sum of terms of the form $\tilde{d}_i \tilde{h}_{\delta k} \alpha_i \beta_i^*$, for some $\tilde{d}_i \in \mathbb{C}$ with $|\tilde{d}_i| < 1$, where $\alpha_i, \beta_i = e_i$ or $h_i$. Hereafter, we use $\tilde{d}_i$ to represent a generic number satisfying $|\tilde{d}_i| < 1$ uniformly on $\mathcal{D}_T(\eta_0, 1)$. Then, we see that $\Delta_G(i, k)$ is a sum of terms of the form
\[
\tilde{d}_i \tilde{g}_{\delta k} G e_{\alpha_i} \beta_i^* \tilde{B}^{(i)} \alpha_i^* R G,
\]
\[
\tilde{d}_i \tilde{g}_{\delta k} G R \tilde{B}^{(i)} \alpha_i^* \beta_i^* G.
\]
\hspace{1cm} (B.2)

Then, the left hand side of the first estimate in (B.1) is a sum of terms of the form
\[
\frac{1}{N} \tilde{d}_i \tilde{g}_{\delta k} X G e_{\alpha_i} \beta_i^* \tilde{B}^{(i)} \alpha_i^* R G e_{\phi},
\]
\[
\frac{1}{N} \tilde{d}_i \tilde{g}_{\delta k} X G R \tilde{B}^{(i)} \alpha_i^* \beta_i^* G e_{\phi}.
\]
\hspace{1cm} (B.3)

By the Cauchy-Schwarz inequality, we have
\[
|\tilde{g}_{\delta k} X G e_{\phi}| \ll \|G e_{\phi}\| = \sqrt{\frac{\text{Im } G_{i i}}{\eta}},
\]
\[
|\beta_i^* \tilde{B}^{(i)} \alpha_i^* R G e_{\phi}| \ll \|G e_{\phi}\| = \frac{\text{Im } G_{i i}}{\eta},
\]
\[
|\tilde{g}_{\delta k} X G R \tilde{B}^{(i)} \alpha_i^* \beta_i^* G| \ll \|G R \tilde{B}^{(i)} \alpha_i^* \beta_i^* G\| = \sqrt{\frac{\text{Im } \tilde{B}^{(i)} \alpha_i^* R G R \tilde{B}^{(i)} \alpha_i^* G_{i i}}{\eta}}.
\]
\hspace{1cm} (B.4)

Note that for $\alpha_i = e_i$,
\[
\alpha_i^* G e_{\phi} = G_{i i}, \quad \alpha_i^* \tilde{B}^{(i)} \alpha_i = b_{i i}^{2} \tilde{g}_{\delta i} g_{i i} = b_{i i}^{2} \tilde{g}_{\delta i},
\]
\hspace{1cm} (B.5)
and for $\alpha_i = h_i$,
\[
\alpha_i^* G_{\alpha i} = \hat{g}_{ii}, \quad \alpha_i^* \tilde{B}^{(i)} R_i G R_i \hat{B}^{(i)} \alpha_i = e_i^* \tilde{B} \hat{G} \hat{B} e_i = \tilde{B}_{ii} - (a_i - z) + (a_i - z) G_{ii}.
\] (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 $O_{\prec}(\Pi_1^2)$. 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
\[
\frac{1}{N} \tilde{d}_i (e_i^* X G \alpha_i)^2 (\beta_i^* \tilde{B}^{(i)} R_i G e_i) (g_i^* X G e_i),
\]
and for $\alpha_i = h_i$,
\[
\alpha_i^* \tilde{B}^{(i)} R_i G R_i \hat{B}^{(i)} \alpha_i = e_i^* \tilde{B} \hat{G} \hat{B} e_i = \tilde{B}_{ii} - (a_i - z) + (a_i - z) G_{ii}.
\] (B.7)
Note that
\[
e_i^* \tilde{B}^{(i)} R_i G e_i = -b_i T_i, \quad h_i^* \tilde{B}^{(i)} R_i G e_i = -(\tilde{B} G)_{ii}.
\]
Hence, we have
\[
|\beta_i^* \tilde{B}^{(i)} R_i G e_i| < 1, \quad |\beta_i^* G e_i| < 1.
\] (B.8)
Further, we claim that
\[
|e_i^* X G \alpha_i|, |e_i^* X G R_i \tilde{B}^{(i)} \alpha_i| \ll \frac{\sqrt{\text{Im} (G_{ii} + \hat{G}_{ii})}}{\eta}.
\] (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 $O_{\prec}(\Pi_1^2)$.

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
\[
\frac{1}{N} \tilde{d}_i (e_i^* X G \alpha_i)^2 (\beta_i^* \tilde{B}^{(i)} R_i G e_i) (g_i^* X G e_i),
\]
and for $\alpha_i = h_i$,
\[
\alpha_i^* \tilde{B}^{(i)} R_i G R_i \hat{B}^{(i)} \alpha_i = e_i^* \tilde{B} \hat{G} \hat{B} e_i = \tilde{B}_{ii} - (a_i - z) + (a_i - z) G_{ii}.
\] (B.10)
Note that
\[
|\beta_i^* \tilde{B}^{(i)} R_i G Q X G \alpha_i| \ll \frac{1}{\eta} \left\| G \alpha_i \right\| \leq \frac{1}{\eta} \sqrt{\frac{\text{Im} (G_{ii} + \hat{G}_{ii})}{\eta}}.
\] (B.11)
Analogously, we have
\[
|\beta_i^* G Q G R_i \tilde{B}^{(i)} \alpha_i| \ll \frac{1}{\eta} \sqrt{\frac{\text{Im} (G_{ii} + \hat{G}_{ii})}{\eta}}.
\] (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 $O_{\prec}(\Psi^2 \Pi_1^2)$. Hence, we obtain the last estimate in (B.1). This concludes the proof of Lemma B.1. \hfill \Box

Proof of Lemma 5.3. The proof is similar to that for Lemma 7.4 in [6]. In the latter, we used $\Psi$ instead of $\Pi_i$ in the statement. However, the proof of Lemma 7.4 in [6] shows readily that the stronger bounds in (5.55) hold for the counterparts of the block additive model (c.f., (7.77), (7.80), (7.81) and (7.87) of [6]). The proof for our additive model given here analogous.

First, by (5.16), (5.17), (5.27), (5.30), and the fact $T_i = T_i - h_i G_{ii}$, we have $|\tilde{S}_i|, |\tilde{T}_i| < 1$, under the assumption (5.12). Then, for the first estimate in (5.55), we have
\[
\frac{1}{N} \sum_k (\tilde{d}_i g_{ik}^{-1} e_k^* X G e_i) = -\frac{1}{2N} - \frac{1}{2N} \sum_k g_{ik} e_k^* X G e_i = -\frac{1}{2N} \left\| g_{ii} \right\| \tilde{h}_i^* X G e_i = O_{\prec}(\frac{1}{N}),
\]
where we used the fact that $\tilde{h}_i^* X G e_i = \tilde{S}_i$ or $\tilde{T}_i$ if $X_i = \tilde{B}^{(i)}$ or $I$, respectively.

Next, we show the second bound in (5.55). It is convenient to set $I^{(i)} := I - e_i^* e_i$. Using (5.37), we get
\[
\frac{1}{N} \sum_k (\tilde{d}_i g_{ik}^{-1} e_k^* X G e_i) = \frac{c_i}{N} e_i^* X G I^{(i)} X e_i + h_i^* \tilde{B}^{(i)} R_i G e_i
\]
\[
+ \frac{c_i}{N} e_i^* X G R_i \tilde{B}^{(i)} I^{(i)} X G e_i (e_i + h_i^*)^* G e_i + \frac{1}{N} \sum_k e_k^* X \Delta_G(i, k) e_i e_k^* X G e_i.
\] (B.13)
The desired estimate of the last term was obtained in the second line of (B.1). Further, using (4.8) we get
\[(e_i + h_i^*) \vec{B}^{(i)} R_i G e_i = -b_i T_i - (\vec{B} G)_{ii} = O_\infty(1), \quad (e_i + h_i)^* G e_i = G_{ii} + T_i = O_\infty(1),\]
where the estimates follow from (5.16) and (5.17). Hence, it suffices to show that
\[
|e_i^* X G I^{(i)} X_i G e_i| \leq \frac{\text{Im} (G_{ii} + \tilde{G}_{ii})}{\eta}, \quad |e_i^* X G R_i \vec{B}^{(i)} I^{(i)} X_i G e_i| \leq \frac{\text{Im} (G_{ii} + \tilde{G}_{ii})}{\eta}.
\]
Note that, by the assumption \(X = I\) or \(A\), both terms in (B.14) can be bounded by
\[C \|G X e_i\| \|G e_i\| = \frac{C}{\eta} \sqrt{\text{Im} (G X X^*)_{ii}} \sqrt{\text{Im} G_{ii}} \leq \frac{C}{\eta} \text{Im} (G_{ii} + \tilde{G}_{ii}).\]
This completes the proof of the second inequality in (5.55). Next, we show the third estimate in (5.55).
In light of the definition of \(T_i\), it suffices to show
\[
\frac{1}{N} \sum_{k}^{(i)} \frac{\partial h_k^*}{\partial g_{ik}} G e_i e_k^* X_i G e_i = O_\infty(1), \quad \frac{1}{N} \sum_{k}^{(i)} h_k^* \frac{\partial G}{\partial g_{ik}} e_i e_k^* X_i G e_i = O_\infty(1),
\]
The first estimate in (B.15) is proved as follows
\[
\frac{1}{N} \sum_{k}^{(i)} \frac{\partial h_k^*}{\partial g_{ik}} G e_i e_k^* X_i G e_i = -\frac{1}{2 \| g_i \|^2} \frac{1}{N} \sum_{k}^{(i)} h_k e_k^* X_i G e_i h_k^* G e_i = -\frac{1}{2 \| g_i \|^2} \frac{1}{N} h_i^* X_i G e_i h_i^* G e_i = O_\infty(1),
\]
where in the last step we again use the fact \(h_i^* \vec{B}^{(i)} G e_i = \tilde{S}_i = O_\infty(1)\) and \(h_i^* G e_i = T_i = O_\infty(1)\). The proof of the second estimate in (B.15) is similar to that of the second inequality in (5.55). It suffices to replace \(e_i^* X\) by \(h_i^*\) 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
\[C \|G h_i\| \|G e_i\| = \frac{C}{\eta} \sqrt{\text{Im} h_i^* G h_i} \sqrt{\text{Im} G_{ii}} = \frac{C}{\eta} \sqrt{\text{Im} G_{ii}} \sqrt{\text{Im} G_{ii}} \leq \frac{C}{\eta} \text{ Im} (G_{ii} + \tilde{G}_{ii}),\]
where we have used (5.43).
Next, we show the fourth estimate in (5.55). Using (5.37) again, we can get
\[
\frac{1}{N} \sum_{k}^{(i)} \text{tr} \left( Q X \frac{\partial G}{\partial g_{ik}} \right) e_k^* X_i G e_i = \frac{c_i}{N^2} (e_i + h_i)^* \vec{B}^{(i)} R_i G Q X G I^{(i)} X_i G e_i
\]
\[+ \frac{c_i}{N^2} (e_i + h_i)^* G Q X G R_i \vec{B}^{(i)} I^{(i)} X_i G e_i + \frac{1}{N} \sum_{k}^{(i)} \text{tr} Q X \Delta_G (i,k) e_k^* X_i G e_i.
\]
The last term above is estimated in (B.1). Using (4.8) and \(\|G\| \leq \eta\), we have
\[
\left| \frac{1}{N^2} (e_i + h_i)^* \vec{B}^{(i)} R_i G Q X G I^{(i)} X_i G e_i \right|
\leq C \frac{1}{N^2 \eta} ((\| G h_i \| + \| G \tilde{B} e_i \|) \| G e_i \| \leq C \frac{1}{N^2 \eta} (\| G h_i \|^2 + \| G \tilde{B} e_i \|^2 + \| G e_i \|^2)
\begin{align}
&= C \frac{1}{N^2 \eta^2} (\text{Im} (h_i^* G h_i + (\vec{B} \tilde{B})_{ii} + G_{ii})) \leq \frac{\text{Im} (G_{ii} + \tilde{G}_{ii})}{N^2 \eta^2}. \quad (B.17)
\end{align}
\]
Here in the last step we again used (5.43) and also fact
\[\text{Im} (\vec{B} \tilde{B}_{ii}) = \eta + \text{Im} ((a_i - z)^2 G_{ii}) = O_\infty(\eta + \text{Im} G_{ii}) = O_\infty(\text{Im} G_{ii}). \quad (B.18)\]
In (B.18), we used (5.8), the first bound in (5.16), and \(\text{Im} G_{ii} \geq \eta\) 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 (5.55) can be proved analogously to the fourth one. The only difference is, instead of the factor \(e_i^* X_i G e_i\) in (6.20), here we have \(e_i^* X_i \tilde{g}_i\) which does not contain any \(G\) factor, which actually makes the estimates even simpler. This completes the proof of Lemma 5.3. \(\square\)
References

[1] Akhiezer, N. I.: *The classical moment problem and some related questions in analysis*, Hafner Publishing Co., New York, 1965.

[2] Ajanki, O., Erdős, L., Krüger, T.: *Universality for general Wigner-type matrices*, arXiv:1506.05098 (2016).

[3] Bao, Z. G., Erdős, L., Schnelli, K.: *Local stability of the free additive convolution*, J. Funct. Anal. **271**(3), 672-719 (2016).

[4] Bao, Z. G., Erdős, L., Schnelli, K.: *Local law of addition of Random Matrices on optimal scale*, Comm. Math. Phys. **349**(3), 947-990 (2017).

[5] Bao, Z. G., Erdős, L., Schnelli, K.: *Convergence rate for spectral distribution of addition of random matrices*, arXiv:1606.03076 (2016).

[6] Bao, Z. G., Erdős, L., Schnelli, K.: *Local single ring theorem on optimal scale*, arXiv:1612.05920 (2016).

[7] Belinschi, S.: *A note on regularity for free convolutions*, Ann. Inst. H. Poincaré Probab. Stat. **42**(5), 635-648 (2006).

[8] Belinschi, S., Bercovici, H.: *A new approach to subordination results in free probability*, J. Anal. Math. **101**(1), 357-365 (2007).

[9] Bercovici, H, Voiculescu, D.: *Free convolution of measures with unbounded support*, Indiana Univ. Math. J. **42**, 733-773 (1993).

[10] Biane, P.: *Process with free increments*, Math. Z. **227**(1), 143-174 (1998).

[11] Bourgade, P., Erdős, L., Yau, H.-T.: *Edge universality of beta ensembles*, Comm. Math. Phys. **332**(1), 261-354 (2014).

[12] Che, Z., Landon, B.: *Local spectral statistics of the addition of random matrices*, arXiv:1701.00513 (2017).

[13] Chistyakov, G. P., Götze, F.: *The arithmetic of distributions in free probability theory*, Cent. Euro. J. Math. **9**, 997-1050 (2011).

[14] Collins, B., Male, C.: *The strong asymptotic freeness of Haar and deterministic matrices*, Ann. Sci. Éc. Norm. Supér. (4) **47**(1), 147-163 (2014).

[15] Diaconis, P., Shahshahani, M.: *The subgroup algorithm for generating uniform random variables*, Probab. Engrg. Inform. Sci. **1**(01), 15-32 (1987).

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

[17] Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: *The local semicircle law for a general class of random matrices*, Electron. J. Probab. **18**(59), 1-58 (2013).

[18] Guionnet, A., Krishnapur, M., Zeitouni, O.: *The single ring theorem*, Ann. of Math. (2) **174**, 1189-1217 (2011).

[19] Kargin, V.: *A concentration inequality and a local law for the sum of two random matrices*, Prob. Theory Related Fields **154**, 677-702 (2012).

[20] Kargin, V.: *Subordination for the sum of two random matrices*, Ann. Proba. **43**(4), 2119-2150 (2015).

[21] Lee, J. O., Schnelli, K.: *Local deformed semicircle law and complete delocalization for Wigner matrices with random potential*, J. Math. Phys. **54**(10), 103504 (2013).

[22] Lee, J. O., Schnelli, K.: *Local law and Tracy–Widom limit for sparse random matrices*, arXiv:1605.08767 (2016).

[23] Mezzadri F.: *How to generate random matrices from the classical compact groups*, Notices Amer. Math. Soc. **54**(5), 592-604 (2007).

[24] Pastur, L., Vasilchuk, V.: *On the law of addition of random matrices*, Comm. Math. Phys. **214**(2), 249-286 (2000).

[25] Tracy, C., Widom, H.: *Level spacing distributions and the Airy kernel*, Comm. Math. Phys. **159**, 151-174 (1994).

[26] Voiculescu, D.: *Addition of certain non-commuting random variables*, J. Funct. Anal. **66**(3), 323-346 (1986).

[27] Voiculescu, D.: *Limit laws for random matrices and free products*, Invent. Math. **104**(1), 201-220 (1991).

[28] Voiculescu, D.: *The analogues of entropy and of Fisher’s information theory in free probability theory, I*, Comm. Math. Phys. **155**, 71-92 (1993).