Tracy-Widom limit for free sum of random matrices

Hong Chang Ji *†    Jaewhi Park ‡§

October 22, 2021

Abstract

We consider fluctuations of the largest eigenvalues of the random matrix model $A + UBU^*$ where $A$ and $B$ are $N \times N$ deterministic Hermitian matrices and $U$ is a Haar-distributed unitary matrix. We prove that the law of the largest eigenvalue weakly converges to the Tracy–Widom distribution, under mild assumptions on $A$ and $B$ to guarantee that the density of states of the model decays as square root around the upper edge. Our proof is based on the comparison of the Green function along the Dyson Brownian motion starting from the matrix $A + UBU^*$ and ending at time $N^{-2/3} \chi$. As a byproduct of our proof, we also prove an optimal local law for the Dyson Brownian motion up to the constant time scale.

AMS Subject Classification (2010): 60B20, 46L54

Keywords: Random matrices, Edge universality, Free additive convolution

October 22, 2021

1 Introduction

In this paper, we consider the sum of two Hermitian random matrices,

$$H = A + UBU^*, \quad (1.1)$$

where $A$ and $B$ are $N \times N$ deterministic Hermitian matrices and $U$ is Haar distributed on the unitary group $U(N)$ of degree $N$. We prove that the law of the largest eigenvalue of the random matrix $H$ in (1.1) converges to the GUE Tracy-Widom distribution $F_2$ with the scale $N^{-2/3}$, under mild assumptions on $A$ and $B$.

The model in (1.1) is one of the most fundamental examples that show the connection between the free probability theory and the space of large Hermitian matrices. The eigenvectors of $A$ and $UBU^*$ are in general position, and thus the two matrices are asymptotically free as the matrix size grows to infinity, as proved by Voiculescu in his influential work [24]. The empirical spectral distribution (ESD) of the ensemble (1.1) converges to the free additive convolution $\mu_\alpha \boxplus \mu_\beta$ where $\mu_\alpha$ and $\mu_\beta$ are limiting eigenvalue distributions of $A$ and $B$, respectively. The convergence of ESD also holds on the local scale [18, 19]. The behavior of the eigenvalues of $H$ on finer scales was studied in [4, 5, 6, 7, 8].

The goal of this paper is to study the law of eigenvalues around the edge on the optimal scale of $N^{-2/3}$. We assume (i) that the averaged eigenvalue distributions $\mu_A$ and $\mu_B$ respectively of $A$ and $B$ converge sufficiently fast to their limits $\mu_\alpha$ and $\mu_\beta$ and (ii) that the densities of $\mu_\alpha$ and $\mu_\beta$ have power-law decay around the upper edge with exponents between $-1$ and 1. The second condition ensures that the

*IST Austria
†Email:hongchang.ji@ist.ac.at
‡Department of Mathematical Sciences, KAIST
§Email:jw-park@kaist.ac.kr
free convolution $\mu_A \boxplus \mu_B$ is regular around the upper edge, that is, its density has square-root decay at the edge; the first condition guarantees that the $N$-dependent convolution $\mu_A \boxplus \mu_B$ inherits the same property.

The most important aspect of our result is that the law of the largest eigenvalue of $H$ not only is independent of $A$ and $B$ but also coincides with that of a GUE. For Gaussian ensembles, Tracy and Widom identified the distribution in [31, 32]; let $\lambda_1$ be the largest eigenvalue of an $(N \times N)$ Gaussian ensemble, normalized so that its spectrum converges to $[-2, 2]$. Then

$$\lim_{N \to \infty} \mathbb{P} \left[ N^{-2/3} (\lambda_1 - 2) < x \right] = F_\beta(x),$$

where $F_\beta(x)$ is a distribution function determined by the parameter $\beta = 1$ or $2$ corresponding respectively to the orthogonal or unitary ensembles. For many random matrix models, the limiting distribution of the largest eigenvalue matches with that of GOE or GUE after proper normalization when the matrix is real symmetric or complex Hermitian, respectively, which is referred to as the edge universality. The edge universality for Wigner matrices was proved in [29, 30, 17, 26], and it was extended to deformed Wigner matrices in [22, 25, 20].

Our proof roughly follows the strategy of [22]. To be specific, we consider a continuous flow $H_t$ of random matrices starting from $H_0 = H$ and for this flow the proof can be divided into three steps. Firstly we prove an optimal local law for $H_t$, secondly we prove the universality for the endpoint of the flow, and lastly we prove that the law of the largest eigenvalue remains the same through the flow. Our choice of the continuous flow is the Dyson Brownian motion (DBM), or Dyson matrix flow,

$$H_t := H + \sqrt{t} W, \quad t \geq 0,$$

where $W$ is a GUE independent of $H$. The most important advantage of taking $t^2$ to be a flow is that we can directly apply the result of Landon and Yau [20], where it was proved that the DBM starting from a regular initial condition reaches its equilibrium at the edge after time $t_0 = N^{-1/3+\delta}$. In other words, the edge statistics of $H_{t_0}$ matches that of GUE, which directly establishes the second step. Here, results of [8] guarantees that the spectrum of the initial matrix $H$ is regular. The shortness of time scale $t_0$ greatly simplifies the last step, compared to [22] where logarithmic time scale was used with Ornstein–Uhlenbeck version of [13].

The bulk of the proof is devoted to the other two components, local law for $H_t$ and comparison over the flow. To prove the local law, we adapt the argument in [8] with some modification to the subordinate system. To summarize, we prove that the resolvent of $H_t$ is subordinate to that of $A$, that is, $G_{H_t}(z) \approx G_A(\omega_{A,t}(z))$ for a complex analytic self-map $\omega_{A,t}$ of the upper half-plane where $G_{H_t}$ and $G_A$ are the resolvents of $H_t$ and $A$, respectively. Here we emphasize that the function $\omega_{A,t}$ captures the dependency on time $t$, and $G_A$ is independent of $t$. Given the local laws, the rest of the proof is devoted to the Green function comparison. To summarize, we prove that the Green function of $H_t$ for spectral parameters around the edge has small enough time derivative, and as a consequence we show that the largest eigenvalue of $H$ and $H_{t_0}$ have the same law. The proof of the Green function theorem mainly concerns expanding the entries of the time derivative of $G_{H_t}$ using Stein’s lemma to express them as a function of entries of $A$ and tracial quantities. Such an argument was first used in [22] for deformed Wigner matrices using Schur complement formula, and later applied to sparse Wigner matrices in [24] with cumulant expansions instead of Schur complements.

In traditional applications of the three step strategy mentioned above, the structure of $H_0$ remained intact along the flow $H_t$ of matrices; for example when $H_0$ is a Wigner matrix, $H_t$ remains a Wigner matrix for all $t$ (see [17, 22, 23] for examples). In the same vein, we may rewrite (1.3) so that it has the same form as $H_0$, namely,

$$H_t = A + (UBU^* + \sqrt{t}W) = A + U_t B_t U_t^*.$$

Here $B_t$ is a diagonal matrix consisting of the eigenvalues of $UBU^* + \sqrt{t}W$, and $U_t$ is a Haar-distributed unitary matrix by the rotational invariance of $W$. Then we might attempt to analyze two flows of matrices $B_t$ and $U_t$, yet we avoid this approach for two reasons. First, for general $B$ the behavior of the spectrum of $B_t$ is much harder to analyze compared to the whole matrix $H_t$. More specifically, while the results of [7] guarantees that $B_t$ has regular edges for $t \sim 1$, the neighborhood on which the square root decay holds true diminishes when $t \ll 1$. In contrast, since $H_0$ has regular edge, the same holds true for $H_t$ no matter how small $t$ is. Second, even though the unitary matrix $U_t$ is Haar distributed for
each fixed $t$, studying it as a stochastic process over $t$ is a difficult task. Indeed these two problems were handled in [10], by assuming stronger conditions on $B$ and introducing a diffusion process $\tilde{U}_t$ on $U(N)$ so that $A + \tilde{U}_t B \tilde{U}_t^*$ has the same local statistics in the bulk as $H_t$ at the time scale $t \gg N^{-1}$. For our purposes it suffices to consider the Dyson matrix flow (1.3) as the sum of three matrices, rather than a perturbation of the sum of two matrices. We explain how we handle the sum of three matrices in the next paragraph.

The major novelty of the proof of local laws for $H_t$ is that we introduce a time dependent, deterministic system of equations (see (3.3)) that characterizes the subordination function $\omega_{A,t}$ above. We take this system as the deterministic equivalent of $G_{H_t}$, which allows us to consider the Brownian motion $\sqrt{t}W$ as one of the leading term, but not a perturbation. In fact, the new system (3.3) is consistent with the one used in [8] in the sense that simply introducing a variant of $F$-transform, the negative reciprocal of Stieltjes transform, can fully reflect the time dependence. This consistency allows us to prove the local law using the exact same strategy as in [8] with some extra bounds for terms originated from the GUE. Due to the same reason, our local laws hold for all finite time scales, which could be of separate interest. Finally, we remark that while the results in [20] can prove an averaged local law for $H_t$, a direct application of their result can neither prove entrywise local laws nor cover constant time scale. To be more specific, along the proof of Green function comparison we need to extract entries of $A$ and $B$ from the Green function of $H_t$, whereas the results of [20] concerns $\mu_A \boxplus \mu_B$ which is an already aggregated quantity.

The proof of Green function comparison for $H_t$ requires new ideas compared to [22] since the randomness in our matrix model originates from Haar unitary matrices. Firstly, as entries of the Haar unitary matrix $U$ are correlated, we use partial randomness decomposition as in [8] to express $U$ in terms of an $N$-dimensional Gaussian vector and a Haar unitary matrix of degree $N - 1$. Secondly, applying Stein’s lemma leads to the derivative of $G$ with respect to the Gaussian vector above, so that traces of matrices of the form $G_{H_t} U B U^* G_{H_{t}}$ emerge. Applying the same argument to these quantities results in more $U$’s, which cannot lead to accessible form. To solve this we use the symmetry of our model, namely to consider $U^* G_{H_t} U$ as the resolvent of

$$H_t := U^* A U + B + \sqrt{t} U^* W U.$$  

Due to the invariance of GUE, we see that the matrix (1.5) in fact has the same form as (1.3). Thus we establish a system of linear equations involving weighted traces of $G_{H_{t}}^2$ and $G_{H_{t}}$, from which the Green function comparison follows. To the best of our knowledge, such calculations involving system of equations did not appear in previous proofs of edge universality.

### 1.1 Related works

The convergence of the ESD of the model in (1.1) was first considered by Voiculescu [34] and extended to a local scale by Kargin [18, 19]. The properties of the free additive convolution in two deterministic measures such as stability, behavior and its qualitative description was studied by Bao, Erdős, and Schnelli [4, 7]. They also established optimal local laws for $H$ and the convergence of the Green function of eigenvalue distribution when the parameter is close to the spectrum $[3] [6] [8]$. As a result, they proved in [8] that the typical eigenvalue spacing is of size $N^{-1}$ and $N^{-2/3}$ around the bulk and edge, respectively. The bulk universality for (1.1) was established by Che and Landon [10], that is, the local eigenvalue statistics of $H$ in the bulk on the scale $N^{-1}$ coincides with that of a GUE.

For the sum of i.i.d. unitarily invariant matrices whose number of summands exceeds certain threshold, the edge universality was obtained by Ahn [1]. As a byproduct, it was proved in [1] that the Tracy–Widom limit holds for (1.1) when $\mu_\alpha$ and $\mu_\beta$ are exactly beta distributions $\text{Beta}(a, b)$ with $a \in [-1/2, \infty)$ and $b \in (-1, 1/2]$.

The linear eigenvalue statistics of (1.1) were studied in [9] [28]. The Gaussian fluctuation of linear eigenvalue statistics on the global scale was obtained by Pastur and Vasilchuk, [25]. On the mesoscopic scale, which concerns relatively few eigenvalues around a fixed energy level, Bao, Schnelli, and Xu proved in [9] a central limit theorem for linear eigenvalue statistics when the energy is in the regular bulk. Also, it is discussed in [9] that the result can be extended to the regular edges up to the scale $N^{-2/3}$, while it is widely regarded as true that the optimal scale at the edge is $N^{-2/3}$. We expect that the analysis in the present paper can shed light on the extension of their result to the full mesoscopic scale at the edge.
The rest of this paper is organized as follows. In Section 2, we present the model and assumptions on it, and rigorously state our main result. In Section 3, we collect preliminary results on free probability, in particular analytic subordination, and recall partial randomness decomposition. Section 4 is devoted to the outline of proof of the main result, where we state the major components of the proof as propositions. In Sections 5 and 6 we prove decoupling lemmas for the resolvent that are used to prove the Green function comparison theorem. Finally in Section 7, we prove the propositions and deduce the Tracy-Widom limit. Appendices A–F mainly concern the proof of local laws for the Dyson matrix flow. Lastly in Appendix G, we present formulas for the derivatives with respect to the Gaussian vector from partial randomness decomposition.

Notational Remark 1.1. \( \mathbb{C} \) denotes the complex open upper half-plane. All quantities, especially matrices and their entries, depend on \( N \) unless otherwise specified. We denote by \( C \) a constant that does not depend on \( N \), which may vary by line.

Notational Remark 1.2. For two sequences \( X := X^{(N)} \) and \( Y := Y^{(N)} \) of random variables with \( Y \geq 0 \), we say that \( Y \) stochastically dominates \( X \) if, for all (small) \( \epsilon > 0 \) and (large) \( D > 0 \),

\[
\mathbb{P} \left[ |X| > N^* Y \right] \leq N^{-D},
\]

for sufficiently large \( N \geq N_0(\epsilon, D) \). In this case we write \( X \prec Y \) or \( X = O(Y) \).

Notational Remark 1.3. We use double brackets to denote integral sets, i.e., \( [n_1, n_2] := [n_1, n_2] \cap \mathbb{Z} \).

Notational Remark 1.4. For \( k \in [1, N] \), we use the shorthand notations

\[
\sum_{i_1, \ldots, i_m := 1}^{N} := \sum_{i_1=1}^{N} \cdots \sum_{i_m=1}^{N}, \quad \sum_{i=1}^{(k)} := \sum_{i=1}^{N}, \quad \text{and} \quad \sum_{i \neq j} := \sum_{i} \sum_{j \neq i}.
\]

2 Definitions and main result

Definition 2.1. Let \( \{a_i : 1 \leq i \leq N\} \) and \( \{b_i : 1 \leq i \leq N\} \) be sequences of real numbers and define \((N \times N)\) diagonal matrices \( A \) and \( B \) by

\[
A := \text{diag}(a_1, \ldots, a_N), \quad B := \text{diag}(b_1, \ldots, b_N).
\]

We denote the empirical spectral distributions of \( A \) and \( B \) by \( \mu_A \) and \( \mu_B \), respectively;

\[
\mu_A := \frac{1}{N} \sum_i \delta_{a_i}, \quad \mu_B := \frac{1}{N} \sum_i \delta_{b_i}.
\]

Let \( U \) be an \((N \times N)\) random unitary matrix drawn from Haar measure on the unitary group \( U(N) \), and we define

\[
H_0 := A + UBU^*.
\]

Definition 2.2. Let \( \mu_\alpha \) and \( \mu_\beta \) be probability measures on \( \mathbb{R} \) satisfying the following:

(i) Both of \( \mu_\alpha \) and \( \mu_\beta \) are compactly supported and absolutely continuous. We denote their densities by \( f_\alpha \) and \( f_\beta \), respectively, and write \([E^-_\alpha, E^+_\alpha]\) and \([E^-_\beta, E^+_\beta]\) for the smallest intervals containing \( \text{supp} \mu_\alpha \) and \( \text{supp} \mu_\beta \), respectively.

(ii) There exist \( t^\alpha_\alpha, t^\beta_\beta \in (-1, 1) \) and positive constants \( \tau_\alpha, \tau_\beta \), and \( C \) such that

\[
C^{-1}(E^+_\alpha - x)^{t^\alpha_\alpha} \leq f_\alpha(x) \leq C(E^+_\alpha - x)^{t^\alpha_\alpha}, \quad \forall x \in [E^+_\alpha - \tau_\alpha, E^+_\alpha], \quad (2.1)
\]

\[
C^{-1}(E^+_\beta - x)^{t^\beta_\beta} \leq f_\beta(x) \leq C(E^+_\beta - x)^{t^\beta_\beta}, \quad \forall x \in [E^+_\beta - \tau_\beta, E^+_\beta]. \quad (2.2)
\]

Assumption 2.3. We assume the following:

(i) For all \( \epsilon > 0 \), we have

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

for sufficiently large \( N \), where \( d_L \) stands for the Lévy distance.
(ii) For all $\delta > 0$, 
\[ \sup \sup \mu_A \leq E_n^+ + \delta, \quad \sup \sup \mu_B \leq E_B^+ + \delta, \]
for sufficiently large $N$.

(iii) There exists $C \in \mathbb{R}$ such that 
\[ \inf \sup \mu_A \geq C, \quad \inf \sup \mu_B \geq C. \]

Under Assumption 2.3, it is well known that the empirical eigenvalue distribution of $H_0$ converges to the free additive convolution $\mu_\alpha \boxplus \mu_\beta$. For our choices of $\mu_\alpha$ and $\mu_\beta$ in Definition 2.2, the measure $\mu_\alpha \boxplus \mu_\beta$ is regular around the upper edge in the sense that it resembles the semi-circle distribution:

**Lemma 2.4** (Theorem 2.2 of [7]). The free convolution $\mu_\alpha \boxplus \mu_\beta$ is compactly supported and absolutely continuous with continuous, bounded density $\rho_0$. Furthermore, there exist $E_{+0} \in \mathbb{R}$ and $\gamma_0 > 0$ such that $E_{+0} = \sup \{x \in \mathbb{R} : \rho_0(x) > 0\}$ and
\[ \lim_{x \nearrow E_{+0}} \frac{\rho_0(x)}{\sqrt{E_{+0} - x}} = \frac{3}{\gamma_0}. \]

Now we are ready to introduce the main result of this paper, the edge universality for $H_0$:

**Theorem 2.5.** Let $A$ and $B$ be deterministic diagonal matrices satisfying Assumption 2.3 and let $\lambda_1$ be the largest eigenvalue of the matrix $H_0$ defined in Definition 2.2. Then we have
\[ \lim_{N \to \infty} \mathbb{P} \left( \gamma_0^{-1} N^{2/3} (\lambda_1 - E_{+0}) \leq s \right) = F_2(s), \]
where $F_2$ is the GUE Tracy-Widom distribution function.

We conclude this section with remarks on Theorem 2.5 and assumptions we imposed.

**Remark 2.6.** Having power-law type decay around the edge as in Definition 2.2 is a typical property among limiting spectral distributions of random matrices. Prime examples are semi-circle (see Definition 3.1) and Marchenko-Pastur distributions, and the arcsine distribution is an example with exponent $t^+ = -1/2$ which is the limit of the ESD of $U + U^*$.

We emphasize that $t^+ < 1$ is not a technical assumption. When either $t^+_\alpha$ or $t^+_\beta > 1$, the density of the free convolution $\mu_\alpha \boxplus \mu_\beta$ may not have square root decay around the edge as in Lemma 2.4. Indeed, it was proved in [24] that when $\mu_\alpha$ is the semi-circle law and $t^+_\beta > 1$, the density of $\mu_\alpha \boxplus \mu_\beta$ can decay as $t^+_\beta$ depending on the ratio of variances of $\mu_\alpha$ and $\mu_\beta$. In this case, the typical eigenvalues spacing of $H$ around the edge would be $N^{-1/(t^+_\alpha + 1)}$, which is incompatible with our result where the largest eigenvalue is scaled by $N^{2/3}$.

**Remark 2.7.** Here we explain the role of Assumption 2.3. Firstly, the assumption (i) guarantees that the $N$-dependent measure $\mu_A \boxplus \mu_B$ inherits the stability of $\mu_\alpha \boxplus \mu_\beta$. We might be able to combine Definition 2.2 with Assumption 2.3 to make a statement on $\mu_A$ and $\mu_B$ that would ensure the stability of $\mu_A \boxplus \mu_B$, uniformly over $N$. While we believe that this is possible following the approach of [1], that is, imposing conditions on the inverse of their Stieltjes transform, we choose the current assumptions to make direct connection with [3] and to avoid using inverse Stieltjes transforms.

The second assumption (ii) should be understood in connection with so-called BBP transition [3], meaning that spiked eigenvalues of $A$ and $B$ may result in those of $H_0$. The corresponding phenomenon was proved in [14], so that when $\max_{i \in [1,N]} a_i$ exceeds certain threshold $\omega_{n,0}^{-1}(E_{+0})$, to be specific, the largest eigenvalue $\lambda_1(H)$ of $H$ converges to a point strictly larger than $E_{+0}$. In this case, we expect that the fluctuation of $\lambda_1(H)$ would be a Gaussian with magnitude $O(N^{-1/2})$ as in [3], which is in a different regime from our result. We will pursue this line of study in a future work. Finally, we mention that while we can generalize our result by allowing few eigenvalues of $A$ to detach from $\sup \mu_\alpha$, but within the threshold above, we refrain ourselves for simplicity.

**Remark 2.8.** The most important and interesting examples would arise when $A$ and $B$ are random. In particular if the optimal local laws as in Theorem 13.1 for $A$ and $B$ are established, eigenvalue rigidity (see Lemma 4.4 for an example) implies $d_L(\mu_A, \mu_B) \ll N^{-1+\epsilon}$ with high probability. Thus our theorem applies to random $A$ and $B$ as long as the optimal local laws are proved and they are unitarily invariant. There are many instances satisfying these criteria, including invariant-ensembles [15] that also cover GUE.
and Wishart ensemble. Other interesting examples concern self-adjoint polynomials of a Haar unitary matrix, such as Re $U^m$ [27].

Also, due to [8], the sum $H$ itself can serve as a summand if the conditions on $A$ and $B$ hold in the whole spectrum (see Corollary 2.8 of [8]). In this case, our result applies to the sum of any finite number of summands, that is, sums of $U_jA_jU_j'$s where $U_j'$s are independent Haar unitary matrices and $A_j'$s satisfy the assumptions of [8, Corollary 2.8].

3 Preliminaries

As mentioned in the introduction, our proof mainly involves the Dyson matrix flow starting from $H_0$, whose ESD is approximated by the free additive convolutions of three measures, $\mu_\alpha \boxplus \mu_\beta \boxplus \mu_{0c}^{(t)}$ or $\mu_\alpha \boxplus \mu_\beta \boxplus \mu_{0c}^{(t)}$. This section collects some complex analytic preliminary results on these free additive convolutions, including the analytic subordination phenomenon.

Definition 3.1. For $t \geq 0$ and a probability measure $\mu$ on $\mathbb{R}$, we define functions $m_\mu, F_{\mu,t}$ on $\mathbb{C}_+$ by

$$m_\mu(z) = \int_{\mathbb{R}} \frac{1}{x-z} d\mu(x), \quad F_{\mu,t}(z) = -\frac{1}{m_{\mu}(z)} + tm_{\mu}(z), \quad z \in \mathbb{C}_+. \quad (3.1)$$

Also we denote the semicircle distribution on $[-2\sqrt{t},2\sqrt{t}]$ by $\mu_{sc}^{(t)}$, that is,

$$d\mu_{sc}^{(t)}(z) = \frac{1}{2\pi t} \sqrt{(4t-x^2)} \, dx.$$

One of the most efficient tools for studying free convolutions, specifically their regularity, is analytic subordination. We present the corresponding result for free additive convolution of two measures and a dilated semi-circle distribution:

Proposition 3.2. Given $t > 0$ and two Borel probability measures $\mu_1$ and $\mu_2$ on $\mathbb{R}$, there exist unique analytic functions $\omega_{1,t}, \omega_{2,t} : \mathbb{C}_+ \rightarrow \mathbb{C}_+$ that satisfy the following:

(i) We have $\text{Im} \omega_{1,t}(z), \text{Im} \omega_{2,t}(z) \geq \text{Im} z$ for all $z \in \mathbb{C}_+$ and

$$\lim_{\eta \searrow \infty} \frac{\omega_{1,t}(i\eta)}{i\eta} = \lim_{\eta \searrow \infty} \frac{\omega_{2,t}(i\eta)}{i\eta} = 1 ;$$

(ii) For all $z \in \mathbb{C}_+$,

$$F_{\mu_1 \boxplus \mu_2 \boxplus \mu_{sc}^{(t)}}(z) = F_{\mu_1,t}(\omega_{1,t}(z)) = F_{\mu_2,t}(\omega_{2,t}(z)) = \omega_{1,t}(z) + \omega_{2,t}(z) - z. \quad (3.2)$$

Proof. The result can be proved with a straightforward modification of Theorem 4.1 of [13], which covers the case of $t = 0$. Or we can apply the same theorem twice, firstly to the free convolution $\mu_1 \boxplus \mu_2$ and then to $(\mu_1 \boxplus \mu_2) \boxplus \mu_{sc}^{(t)}$. The second proof also reveals the relation

$$\omega_{1,t}(z) = \omega_{1,0}(z + tm_\rho(z)), \quad \omega_{2,t}(z) = \omega_{2,0}(z + tm_\rho(z)),$$

where we abbreviated $\rho = \mu_1 \boxplus \mu_2 \boxplus \mu_{sc}^{(t)}$. $\square$

For simplicity, we use the following abbreviations:

$$m_{\mu_\alpha} \equiv m_\alpha, \quad m_{\mu_\beta} \equiv m_\beta, \quad m_{\mu_\alpha} \equiv m_\alpha, \quad m_{\mu_B} \equiv m_B,$$

$$F_{\mu_\alpha,t} \equiv F_{\alpha,t}, \quad F_{\mu_\beta,t} \equiv F_{\beta,t}, \quad F_{\mu_\alpha,t} \equiv F_{\alpha,t}, \quad F_{\mu_B,t} \equiv F_{B,t},$$

$$\mu_\alpha \boxplus \mu_\beta \boxplus \mu_{sc}^{(t)} \equiv \mu_t, \quad \mu_\alpha \boxplus \mu_\beta \boxplus \mu_{sc}^{(t)} \equiv \mu_t.$$

We further denote the subordination functions corresponding to the pairs $(\mu_\alpha, \mu_\beta)$ and $(\mu_A, \mu_B)$ respectively by $(\omega_{\alpha,t}, \omega_{\beta,t})$ and $(\omega_{\alpha,t}, \omega_{\beta,t})$, so that

$$F_{\alpha,t}(\omega_{\alpha,t}(z)) = F_{\beta,t}(\omega_{\beta,t}(z)) = F_{\mu_\alpha}(z), \quad F_{A,t}(\omega_{A,t}(z)) = F_{B,t}(\omega_{B,t}(z)) = F_{\mu_B}(z).$$
Lemma 3.3. The boundary behaviors of the subordination functions are studied by Belinschi in the series of papers [10, 11, 12]. In particular it is proved that if $\mu_1(a) + \mu_2(b) < 1$ for all $a, b \in \mathbb{R}$, then the corresponding subordination functions $\omega_{1,0}$ and $\omega_{2,0}$ extend continuously to $\mathbb{C}_+ \cup \mathbb{R}$, possibly attaining value $\infty$. We can easily see that all measures considered in the present paper satisfy the assumption. In particular, $\omega_{\alpha,t}, \omega_{\beta,t}, \omega_{A,t}$ and $\omega_{B,t}$ extend continuously to $\mathbb{C}_+ \cup \mathbb{R}$ for each fixed $t \geq 0$.

We denote the upper edges of $\mu_t$ and $\tilde{\mu}_t$ respectively by $E_{+,t}$ and $\tilde{E}_{+,t}$, and consider the spectral domain

$$D_\tau(\eta_1, \eta_2) \equiv D_{\tau,t}(\eta_1, \eta_2) := \{ z = E + i\eta \in \mathbb{C}_+ : E \in [E_{+,t} - \tau, \tau^{-1}], \eta \in (\eta_1, \eta_2) \},$$

for $\tau, \eta_1, \eta_2, t \geq 0$. Furthermore we denote $\kappa \equiv \kappa_t(z) = |z - E_{+,t}|$ for $z \in \mathbb{C}_+$.

In the following lemma, we present our results on the behavior of $\omega_{\alpha,t}(z)$ and $\omega_{\beta,t}(z)$. Its proof is deferred to Appendix A.

**Lemma 3.3.** Let $\omega_{\alpha,t}, \omega_{\beta,t}$ be the subordination functions corresponding to the pair $(\mu_\alpha, \mu_\beta)$. Then the following hold true:

(i) The functions $(t, z) \mapsto \omega_{\alpha,t}(z), \omega_{\beta,t}(z)$, and $m_{\mu_t}(z)$ are continuous on $[0, \infty) \times (\mathbb{C}_+ \cup \mathbb{R})$.

(ii) There exists a (small) constant $\tau > 0$ such that for all fixed $\eta_M > 0$ we have

$$\sup_{t \in [0,1]} \sup_{z \in D_\tau(0, \eta_M)} |\omega_{\alpha,t}(z)| + |\omega_{\beta,t}(z)| \leq C.$$

(iii) There exist constants $\tau > 0$ and $k_0$ such that

$$\inf \{ |\omega_{\alpha,t}(z) - x| : t \in [0,1], z \in D_\tau(0, \infty), x \in \text{supp} \mu_\alpha \} \geq k_0,$$

$$\inf \{ |\omega_{\beta,t}(z) - x| : t \in [0,1], z \in D_\tau(0, \infty), x \in \text{supp} \mu_\beta \} \geq k_0.$$

Furthermore, $\omega_{\alpha,t}(E_{+,t}) > E_{+,t}^{\uparrow}$ and $\omega_{\beta,t}(E_{+,t}) > E_{+,t}^{\uparrow}$.

(iv) For each $t \geq 0$, the edge $E_{+,t}$ satisfies the following equation.

$$(F'_\alpha(\omega_{\alpha,t}(E_{+,t})) - 1)(F'_\beta(\omega_{\beta,t}(E_{+,t})) - 1) - 1 = 0.$$  

(v) For each $t \geq 0$, the measure $\mu_t$ has a continuous density $\rho_t$ around $E_{+,t}$ and there exists a constant $\gamma_t > 0$ such that the following holds:

$$\lim_{x \nearrow E_{+,t}} \frac{\rho_t(x)}{\sqrt{E_{+,t} - x}} = \frac{\gamma_t^{3/2}}{E_{+,t}}.$$  

Furthermore, $\gamma_t \sim 1$ and $\frac{d}{dt} \gamma_t \sim \sqrt{t}$ for $t \in [0,1]$.

3.1 Partial randomness decomposition

We conclude this section by introducing the partial randomness decomposition of a Haar unitary matrix and notations associated to it. They will be extensively used throughout the remaining sections.

Recall that $U$ is an $(N \times N)$ Haar unitary random matrix. For all $i \in \llbracket 1, N \rrbracket$, define

$$v_i := U e_i, \quad \theta_i := \arg e_i \cdot v_i, \quad h_i := e^{-\theta_i} v_i, \quad \hat{h}_i := h_i - h_i e_i, \quad \ell_i := \frac{\sqrt{2}}{\| e_i + h_i \|}$$

Then the partial randomness decomposition of $U$ stands for the following factorization of $U$:

$$U^{(i)} := -e^{-i\theta_i} R_i U, \quad \text{where} \quad R_i := I - \tau_i r_i, \quad r_i := \ell_i (e_i + h_i).$$

We further define $\tilde{B}^{(i)} := U^{(i)} B U^{(i)*}$. Then the following hold true;
• The matrix $R_i$ is a Householder reflection, that is, $R_i = R_i^* = R_i^2 = I$.
• The $i$-th row and column vectors of $U^{(i)}$ are $e_i^*$ and $e_i$, respectively.
• The $(i, i)$-matrix minor of $U^{(i)}$ is independent of $v_i$ and Haar distributed on $U(N - 1)$.
• $R_i e_i = -h_i$ and $R_i h_i = -e_i$, so that $e_i^* B = -h_i^* \tilde{B}^{(i)} R_i$ and $h_i^* \tilde{B} = b_i h_i^* = -e_i \tilde{B}^{(i)} R_i$.

Since $v_i$ is uniformly distributed on the unit sphere $S^N_0$, we may take a Gaussian vector $\tilde{g}_i \sim N(0, N^{-1} 1_N)$ such that $v_i = \tilde{g}_i/\|\tilde{g}_i\|$. Using the Gaussian vector, we also define

$$g_i := e_i^* \tilde{g}_i \quad \text{and} \quad \tilde{g}_i := g_i - g_i e_i.$$ 

Note that $h_i = g_i/\|g_i\|$, $g_{ij} \sim N(0, N^{-1})$ for $i \neq j$, and $g_i$ is independent of $U^{(i)}$.

### 4 Outline of the proof

The main idea of our proof is to apply Green function comparison to the Dyson matrix flow $H_t$ starting from $H_0$, whose precise definition is as follows:

**Definition 4.1.** For $t \geq 0$, we define the $(N \times N)$ random matrix $H_t$ as

$$H_t = H_0 + \sqrt{t} W, \quad X_t = \gamma_t H_t,$$  

(4.1)

where $W$ is a GUE independent of $H_0$ and $\gamma_t$ is defined in \[3.6\]. Also we define $t_0 = N^{-1/3+\chi}$ where $\chi > 0$ is a sufficiently small constant. For $z \in \mathbb{C}$, we define the resolvent and Green function of $X_t$ as

$$G_t := (\gamma_t H_t - z)^{-1}, \quad m(z) \equiv m_t(z) := \text{tr} G_t.$$

We define their symmetric analogues as follows;

$$W_t = U^* W U, \quad \mathcal{H}_t = H_0 + \sqrt{t} W, \quad \mathcal{X}_t = \gamma_t \mathcal{H}_t, \quad \mathcal{G}_t = (\gamma_t \mathcal{H}_t - z)^{-1}.$$  

(4.2)

Note that the density of limiting eigenvalue distribution of $X_t$ decays as $\frac{1}{\sqrt{N}} \sqrt{\gamma_t E_{+t} - x}$ due to the scaling factor $\gamma_t$. For simplicity, we often omit the subscript $t$ to denote $H \equiv \mathcal{H}_t$, $G \equiv \mathcal{G}_t$, et cetera.

### 4.1 Proof of Theorem 2.5

In order to compare the largest eigenvalues of $H_0$ and $H_t$, we employ the Green function comparison whose precise statement is as follows.

**Proposition 4.2** (Green function comparison). Let $\epsilon > 0$ and $\eta_0 = N^{-2/3-\epsilon}$. Let $E_1, E_2 \in \mathbb{R}$ with $|E_1|, |E_2| \leq N^{-2/3+\epsilon}$ and $F : \mathbb{R} \to \mathbb{R}$ be a smooth function satisfying

$$\sup_{x \in \mathbb{R}} |F^{(\ell)}(|x| + 1)^{-C}| \leq C, \quad \ell = 1, 2, 3, 4.$$ 

Then there exists a constant $C' > 0$ such that the following hold for all sufficiently large $N$ and sufficiently small $\epsilon' > 0$;

$$\left| \mathbb{E}[F \left( \int_{E_1}^{E_2} \text{Im} \text{Tr} G_0(L_{+0} + E + i\eta_0) dE \right) - F \left( \int_{E_1}^{E_2} \text{Im} \text{Tr} G_{t_0}(L_{+0} + E + i\eta_0) dE \right) \right] \right| \leq N^{-1/3+C'\epsilon'},$$  

(4.3)

where $L_{+t} := \gamma_t E_{+t}$.

In the next proposition, we prove that the Tracy-Widom limit holds for $H_{t_0}$.

**Proposition 4.3.** Let $F : \mathbb{R}^k \to \mathbb{R}$ be a smooth function such that $\|F\|_\infty \leq C$ and $\|\nabla F\|_\infty \leq C$ for some $C > 0$. Then

$$\left| \mathbb{E}[F(\eta_0 N^{2/3}(\lambda_1, t_0 - E_{+t_0}), \cdots, \eta_0 N^{2/3}(\lambda_k, t_0 - E_{+t_0}))] \right. \left. - E^{\text{GUE}}[F(N^{2/3}(\mu_1 - 2), \cdots, N^{2/3}(\mu_k - 2))] \right| \leq N^{-c}$$  

(4.4)

holds with high probability for some $c > 0$, where $\lambda_1, \cdots, \lambda_k$ are the eigenvalues of $H_{t_0}$.
We postpone the proofs of Propositions 4.2 and 4.3 to Section 4.
Along the proofs of the above propositions, we need the following local laws for $H_t$ for $t \in [0, t_0]$ near the edge:

**Proposition 4.4.** Under the notations of Proposition 4.2, the followings hold uniformly over $t \in [0, t_0]$ and $z \in \{E + L_{t, t} + i\eta : E \in [E_1, E_2]\}$:

\[
\max_{a, b} \left| \gamma_{i}G_{ab} - \delta_{ab} - \frac{1}{a_{a} - \omega_{a, t}(E_{i, t})} \right| + \left| (U^{*} G)_{ab} \right| \lesssim N^{-1/3 + \epsilon},
\]

\[
\max_{a, b} \left| \gamma_{j}G_{ab} - \delta_{ab} - \frac{1}{b_{b} - \omega_{b, t}(E_{j, t})} \right| + \left| (G U)_{ab} \right| \lesssim N^{-1/3 + \epsilon},
\]

\[
\max_{a, b} \left| \gamma_{i}G_{ab} - \delta_{ab} - \frac{1}{a_{a} - \omega_{a, t}(E_{i, t})} \right| + \left| (U^{*} G)_{ab} \right| \lesssim N^{-1/3 + \epsilon},
\]

\[
\max_{a, b} \left| \gamma_{j}G_{ab} - \delta_{ab} - \frac{1}{b_{b} - \omega_{b, t}(E_{j, t})} \right| + \left| (G U)_{ab} \right| \lesssim N^{-1/3 + \epsilon},
\]

\[
\max_{1 \leq i \leq cN} \frac{1}{N} |\lambda_{i, t} - \gamma_{i, t}| \lesssim N^{-\frac{1}{2}}, \tag{4.9}
\]

where $\gamma_{j, t}$ is the smallest real number such that

\[
\mu_{i}((\infty, \gamma_{j, t}]) = \mu_{i} \sqcup \mu_{\beta} \sqcup \mu_{\alpha}((\infty, \gamma_{j, t}]) \geq \frac{N - j + 1}{N}.
\]

The proofs of Proposition 4.4 and Lemma 4.5 are presented in Appendix B. Armed with Propositions 4.2, 4.3, 4.4 and Lemma 4.5, we now prove Theorem 2.5.

**Proof of Theorem 2.5.** We follow the arguments in the proof of Theorem 2.10 in [24] with some modifications. Namely, we can simply plug the proof components above into their counterparts in [24]. Below we briefly explain the role of each component.

From the rigidity of eigenvalues of $H_t$, Lemma 4.5, and applying Lemma 3.3, we can approximate the cumulative distribution function of $\lambda_{1, t}$ via local estimate of the Green function $G_t$ as in Proposition 7.1 of [24]. In other words, the cumulative distribution function of $\lambda_{1, t}$ and $\lambda_{1, t_0}$ can be approximated by the first and the second term on the left side of (4.3), respectively. Hence Proposition 4.2 guarantees that the cumulative distribution functions of $\lambda_{1, 0}$ and of $\lambda_{1, t_0}$ have the same limit. Since Proposition 4.3 shows that the distribution of $\lambda_{1, t_0}$ weakly converges to the Tracy-Widom distribution $F_2$, Theorem 2.5 follows.

**4.2 Sketch of the proof of Proposition 4.2**

In this section we give the sketch of the proof of Proposition 4.2. We first collect two technical tools used throughout the remaining sections. The first tool is Stein’s lemma; for a continuous function $F : \mathbb{C} \to \mathbb{C}$ with bounded derivatives and a standard complex Gaussian random variable $X$, that is, $\text{Re } X$ and $\text{Im } X$ are i.i.d. with law $\mathcal{N}(0, 1/2)$, we have

\[
\mathbb{E}[X F(X)] = \mathbb{E}\left[\frac{\partial}{\partial X} F(X)\right], \quad \mathbb{E}[X F(X)] = \mathbb{E}\left[\frac{\partial}{\partial X} F(X)\right].
\]

Here $\frac{\partial}{\partial X}$ and $\frac{\partial}{\partial X}$ denote the holomorphic and anti-holomorphic derivatives, i.e. $\frac{\partial}{\partial X} = \frac{1}{2} \left( \frac{\partial}{\partial \text{Re } X} - i \frac{\partial}{\partial \text{Im } X} \right)$ and $\frac{\partial}{\partial X} = \frac{1}{2} \left( \frac{\partial}{\partial \text{Re } X} + i \frac{\partial}{\partial \text{Im } X} \right)$.

Next, for readers’ convenience, we present below formulas for partial derivatives of the resolvent $G_t = (X - z)^{-1}$ with respect to the Gaussian vector $g$ introduced in Section 3.1. These formulas can be proved using the fact that $R_i$ is a Householder reflection and that $U^{(i)}$ is independent of $g$. We omit the proof for it is purely computational.
Lemma 4.6. For $b \neq d \in [1, N]$, we have

$$
\frac{\partial G}{\partial g_{bd}} = -\frac{\ell^2}{\|g_b\|} G[e_d(e_b + h_b)^*], \quad \frac{\partial G}{\partial g_{bd}} = \frac{\ell^2}{2\|g_b\|} h_{bd} G[(e_b + 2h_b)e_b^*], \quad \frac{\partial G}{\partial g_{bd}}. 
$$

Similarly, we have

$$
\frac{\partial G}{\partial g_{bd}} = -\frac{\ell^2}{\|g_b\|} G[\tilde{B}, (e_b + h_b)e_b^*] \quad \frac{\partial G}{\partial g_{bd}} = \frac{\ell^2}{2\|g_b\|} h_{bd} G[\tilde{B}, (e_b + 2h_b)^*]. 
$$

Now we sketch the proof of Proposition 4.2. Define

$$
Y \equiv Y_t := N \int_{E_1}^{E_2} \text{Im} \, m_i(E + L_+ (t) + i\eta_0) \, dE.
$$

To prove the proposition, it is enough to show that the time derivative of $\mathbb{E}[F(Y_t)]$ is small. Differentiating $F(Y_t)$ with respect to $t$, we obtain

$$
\frac{d\mathbb{E}[F(Y_t)]}{dt} = \mathbb{E}[F'(Y_t) \frac{dY}{dt}] = \mathbb{E}[F'(Y_t) \text{Im} \int_{E_1}^{E_2} \sum_a \frac{dG(L_+ + E + i\eta_0)_{aa}}{dt} \, dE]
$$

After taking the integral over $E_1$ to get

$$
\sum_{b,c} \mathbb{E}[F''(Y_t)W_{ba}G_{ac}G_{cb}] = \frac{1}{N} \gamma \sqrt{t} \sum_{b,c} \mathbb{E}[F''(Y_t)G_{ab}G_{ac}G_{cb}].
$$

In summary, we have

$$
\frac{d\mathbb{E}[F(Y_t)]}{dt} = \mathbb{E}[F'(Y_t) \text{Im} \int_{E_1}^{E_2} \left( L_+ \text{Tr} G^2 - \gamma \text{Tr} GHG \right) dE]
$$

Note that $\frac{\partial Y}{\partial X_{bc}}$ in the last term of (4.11) is not integrated over $E$.

In the first step of the proof of Proposition 4.2, we decouple the index $a$ from the resolvent entries in (4.11). More specifically, in Proposition 4.2, we express the $m$ in terms of deterministic functions of the index $a$ and weighted traces of $G^2$. After that, in Section 7.1, we observe that the all terms cancel except an error of size $O(N^{1/6+3\varepsilon})$, whose contribution would be $O(N^{-1/6+3\varepsilon})$ after taking the integral over $t$.

5 Decoupling lemmas

In this section, we expand $\mathbb{E}[F'(Y_t)(\tilde{B}G^2)_{aa}]$ in order to decouple the index $a$ from it. Using the fact that $zG = \gamma HG - I$, we have

$$
z\mathbb{E}[F'(Y_t)(G^2)_{aa}] = \gamma \mathbb{E}[F'(Y_t)(a_0(G^2)_{aa} + (\tilde{B}G^2)_{aa} + \sqrt{t}(WG^2)_{aa})] - \mathbb{E}[F'(Y_t)G_{aa}],
$$

so that it directly relates to the decoupling lemma for $(G^2)_{aa}$.

We will see that the expansion of $\mathbb{E}[F'(Y_t)(\tilde{B}G^2)_{aa}]$ involves $\mathbb{E}[F'(Y_t)h_{sa}^*G_{sa}]$ and vice versa. Our final result then follows from combining the two expansions. Along the expansion, we often have to deal with the quantities of the form

$$
\mathbb{E}[F'(Y_t)x(\tilde{B}G)_{ab}G_{ba}] \quad \text{or} \quad \mathbb{E}[F'(Y_t)\gamma h_{sa}^*G_{sa}],
$$
where $x$ is a (weighted sum of) diagonal entries of $G$ or $G$. In this case we seek to replace the factor $x$ by its deterministic equivalent, in order to derive a deterministic system of linear equations for $\mathbb{E}[F(Y)(BG^2)_{aa}]$ and $\mathbb{E}[F(Y)h_{12}G^2e_a]$. To this end, we introduce the following notations for $k = 1, 2, 3$:

$$D := \text{diag} \left( \gamma \frac{\alpha - \alpha_0}{\alpha_0} \right)_{a \in [1,N]}, \quad \mathcal{D} := \text{diag} \left( \gamma \frac{b_k - \beta}{\beta} \right)_{b \in [1,N]},$$

$$d_k := \text{tr} (A^{k-1}D), \quad \varrho_k := \text{tr} (B^{k-1}D),$$

$$(\varrho) := \omega_\beta (\omega_\beta + \hat{m}) \mathcal{d}_1 = (2\omega_\beta + \hat{m} - 1) \mathcal{d} + \varrho_3,$$

where we defined $\hat{\omega}_\alpha := \omega_\alpha (E_k), \hat{\omega}_\beta := \omega_\beta (E_\alpha), \text{ and } \hat{m} = \gamma^{-1} m_\mu (E_\alpha)$. Note that Proposition 4.3 implies $d_k, \varrho_k$, and all entries of $D, \mathcal{D}$ are $O(N^{-1/3+\epsilon})$. Furthermore, we have

$$\mathcal{Y} = (\varrho_\beta \hat{m} + 1) + \varrho_3 - (\varrho_\beta \hat{m} + 1 + \varrho_2)^2 + (\hat{m} + \varrho_1) (\omega_\beta \hat{m} + 1) + \mathcal{d}_3 = (\varrho) + O(N^{-2/3+2\epsilon}),$$

where $\mathcal{Y}$ is defined in (5.3) and we used the fact that $\gamma \text{tr} \hat{B}^{k-1}G - \varrho_k = O(1)$, which can be proved combining

$$\gamma \text{tr} \hat{B}^{k-1}G - \varrho_k = \frac{1}{N} \sum_k \varrho_k b_k^{k-1} = O(\varrho),$$

with Lemma 3.3 Thus, (5.3) and (5.4) imply that $(\varrho) = O(N^{-1/2+\epsilon})$.

To simplify presentation, we introduce three dimensional complex vectors $\mathbf{x}_a, x_a$ and $(3 \times 3)$ matrices $Z_a, Z_a, \mathcal{Z}_a$, and $\hat{Z}_a$ for each $a \in [1, N]$ defined as follows:

$$x_{ai} := \mathbb{E}[F(Y)(G\bar{A}^{k-1}g)_{aa}], \quad Z_{ak} := \mathbb{E}[F(Y) d_k (G\bar{B}^{k-1}G)_{aa}], \quad \mathcal{Z}_{ak} := \mathbb{E}[F(Y) d_k (G\bar{A}^{k-1}g)_{aa}],$$

$$x_{ai} := \mathbb{E}[F(Y)(G\bar{B}^{k-1}G)_{aa}], \quad Z_{ak} := \mathbb{E}[F(Y) d_k (G\bar{A}^{k-1}g)_{aa}], \quad \mathcal{Z}_{ak} := \mathbb{E}[F(Y) d_k (G\bar{B}^{k-1}G)_{aa}],$$

for $i, k, l = 1, 2, 3$. The same vectors and matrices without subscript $a$ stand for sums over $a$ (for examples, $\mathbf{x} := \sum_a x_a$ and $Z = \sum_a Z_a$). Finally, we define

$$u_\alpha := \left( \int_{(x - \hat{\omega}_\alpha)} d\mu_\alpha (x) \quad \int_{(x - \hat{\omega}_\alpha)} d\mu_\alpha (x) \quad \int_{(x - \hat{\omega}_\alpha)} d\mu_\alpha (x) \right)^\top,$$

$$v_\alpha := \left( \frac{1}{(x - \hat{\omega}_\alpha)^2} - \frac{\hat{m}}{(x - \hat{\omega}_\alpha)^2} - 2\hat{\omega}_\alpha \right),$$

and $u_\beta, v_\beta$ are defined by the same equation with roles of $\mu_\alpha$ and $\mu_\beta$ interchanged. By Lemma 3.3 all components of $u_\alpha, u_\beta, v_\alpha, v_\beta$ are bounded and the first components of $u_\alpha$ and $u_\beta$ are bounded below.

Now we state the main result of this section, in which we decouple the index $a$ from $\mathbb{E}[F(Y)(BG^2)_{aa}]$.

**Proposition 5.1.** Under the settings in Proposition 4.3, the following holds true uniformly over $z \in \{E + L_0 + i\mathbb{R} : E \in [E_1, E_2]\}$:

$$\mathbb{E}[F(Y)(BG^2)_{aa}] = (\hat{m} + \hat{\omega}_\beta) x_{ai} - \frac{1}{N(a_0 - \hat{\omega}_\alpha)} v_{\beta}^\top x + \frac{(v_{\beta}^\top u_\beta)^2}{u_{\beta}^2} \left( - \frac{1}{N(a_0 - \hat{\omega}_\alpha)^2} + \frac{\hat{m}}{N(a_0 - \hat{\omega}_\alpha)} \right) \left( 2Z_{11} + \frac{\gamma}{N} \int_{E_1}^{E_2} \mathbb{E}[F'(Y) \text{Tr} G^2 \text{Im}(G^2)] dE \right) + O(N^{-1/6+2\epsilon}).$$

Along the proof of Proposition 5.1 we repeatedly apply Stein’s lemma with respect to the Gaussian random vector $g_\beta$. The following lemma shows the result of the first application.
Lemma 5.2. Under the conditions in Proposition 5.1, we have

\[
\mathbb{E}[F'(Y)(\vec{B}G^2)_{aa}] = (\tilde{m}^{-1} + \tilde{\omega}_\beta x_{a1} + \mathbb{E}[F'(Y) \left(- (\tilde{\omega}_\beta + \tilde{m}^{-1}) d_1 + d_2\right) (\vec{B}G^2)_{aa}] + \mathbb{E}[F'(Y) \left((b_a - \tilde{\omega}_\beta) d_2 - (b_a - \tilde{\omega}_\beta)(\tilde{m}^{-1} + \tilde{\omega}_\beta)d_1 - (b)\right) h_a G^2 e_a] + \mathbb{E}[F'(Y) \left(- d_3 + (\tilde{\omega}_\beta + \tilde{m}^{-1}) d_2\right) (G^2)_{aa}]
\]

\[
+ \gamma \mathbb{E}[F'(Y) (\nabla \vec{B}G^2 - (\tilde{\omega}_\beta + \tilde{m}^{-1}) \nabla T^2) \frac{(\vec{B}G^2)_{aa}}{N}]
\]

\[
+ \gamma \mathbb{E}[F'(Y) \left(- \nabla \vec{B}G^2 + (\tilde{\omega}_\beta + \tilde{m}^{-1}) \nabla T^2\right) \frac{G_{aa}}{N}]
\]

\[
+ \gamma \mathbb{E}[F'(Y) \left(-b_a(\tilde{\omega}_\beta + \tilde{m}^{-1}) \nabla T^2 + (b_a + \tilde{\omega}_\beta + \tilde{m}^{-1}) \nabla \vec{B}G^2\right) \frac{h_a^2 G e_a}{N}]
\]

\[
+ \frac{1}{N} \sum_{e} \mathbb{E}[F''(Y) \frac{\partial Y}{\partial g_{ac}} \left(- (\vec{B}G^2)_{ca} + (\tilde{\omega}_\beta + \tilde{m}^{-1})(G^2)_{ca}\right)] + O(N^{-1/6+2\epsilon}).
\]

Proof. The proof consists of expansions of two quantities: \(\mathbb{E}[F'(Y)(\vec{B}G)_{ab} G_{ba}]\) and \(\mathbb{E}[F'(Y)h^a_e G e_a G_{ba}]\).

Step 1: Expansion of \(\mathbb{E}[F'(Y)(\vec{B}G)_{ab} G_{ba}]\)

We first write

\[
\mathbb{E}[F'(Y)(\vec{B}G)_{ab} G_{ba}] = - \mathbb{E}[F'(Y)h^a_e \vec{B}^{(a)}_e R_{a} G e_a G_{ba}]
\]

\[
= - \sum_c \mathbb{E}[F'(Y) \frac{\partial}{\partial g_{ac}} e^c_r \vec{B}^{(a)}_e R_{a} G e_a G_{ba}]
\]

\[
= - \mathbb{E}[F'(Y) \frac{\partial}{\partial g_{ac}} e^c_r h^a_e G e_a G_{ba}] - \frac{1}{N} \sum_{c \neq a} \mathbb{E}[\frac{\partial}{\partial g_{ac}} \left(F'(Y) \|g_a\|^{-1} e^c_r \vec{B}^{(a)}_e R_{a} G e_a G_{ba}\right)]
\]

\[
= - \frac{1}{N} \sum_{c \neq a} \mathbb{E}[\frac{\partial}{\partial g_{ac}} \left(F'(Y) \|g_a\|^{-1} e^c_r \vec{B}^{(a)}_e R_{a} G e_a G_{ba}\right)] + O(N^{-7/6+2\epsilon}),
\]

where we used in the last equality that

\[
g_{aa} \propto N^{-1/2}, \quad \|g_a\| \sim 1, \quad h^a_e G e_a \propto N^{-1+3\epsilon}, \quad G_{ba} \propto N^{-1+3\epsilon}.
\]

We calculate the partial derivative on the right-hand side of (5.8) as

\[
\frac{\partial}{\partial g_{ac}} \left(F'(Y) \|g_a\|^{-1} e^c_r \vec{B}^{(a)}_e R_{a} G e_a G_{ba}\right)
\]

\[
= -\gamma F'(Y) \|g_a\|^2 e^c_r \vec{B}^{(a)}_e R_{a} G \left[ (e_c (e_a + h_a)^* , \vec{B} G e_a e^*_c + e_b e^*_c G (e_c (e_a + h_a)^* , \vec{B}) \right] G e_a
\]

\[
+ F''(Y) \|g_a\|^{-1} \frac{\partial Y}{\partial g_{ac}} e^c_r \vec{B}^{(a)}_e R_{a} G e_a G_{ba} + F'(Y) \|g_a\|^{-1} e^c_r \vec{B}^{(a)}_e R_{a} \left(\Delta_G(a,c) e_b G_{ba} + G_{ba} \Delta_G(a,c) b_{ba}\right)
\]

\[
+ F'(Y) \|g_a\|^{-1} e^c_r \vec{B}^{(a)}_e \frac{\partial R_a}{\partial g_{ac}} G e_a G_{ba} + \frac{\partial \|g_a\|^{-1} e^c_r \vec{B}^{(a)}_e R_{a} G e_a G_{ba}},
\]

where we denoted

\[
\Delta_G(a,c) := \gamma \frac{\|g_a\|^2}{2} \vec{B}_c G (2 h_b + e_b) e^*_c \vec{B} G + \gamma \frac{\|g_a\|^2}{2} \vec{B} \Delta_G(a,c) e_b G_{ba} + \frac{\partial \|g_a\|^{-1} e^c_r \vec{B}^{(a)}_e R_{a} G e_a G_{ba}},
\]

We now consider the contribution of the first term on the right-hand side of (5.9). First of all, to convert the sum \(\sum_{c \neq a}\) into \(\sum_{c}\), we plug in \(c = a\) into the first term of (5.9):

\[
- \frac{1}{N} e^c_r \vec{B}^{(a)}_e R_{a} G (e_a + h_a)^* , \vec{B} G e_a G_{ba} = \frac{1}{N} \vec{B} \|h_a^c G (e_a + h_a)^* , \vec{B} G e_a G_{ba} = O(N^{-2+3\epsilon}).
\]
Thus we have
\[
\gamma \frac{1}{N} \sum_{c \neq a} \mathbb{E}[F'(Y) \frac{\ell^2}{\|g_a\|^2} e^{c} \tilde{B}^{(a)} R_a G[e_c(e_a + h_a)^*, \tilde{B}]G_{e_bG_{ba}}] \\
= \mathbb{E}[F'(Y) \text{tr}(\tilde{B}G)\tilde{B}G_{ab}G_{ba}] + \mathbb{E}[F'(Y)h_a \text{tr}(\tilde{B}G)h^*_{ab}G_{ebG_{ba}}] \\
- \mathbb{E}[F'(Y) \text{tr}(\tilde{B}G\tilde{B})G_{ab}G_{ba}] - \mathbb{E}[F'(Y) \text{tr}(\tilde{B}G\tilde{B})h^*_{ab}G_{ebG_{ba}}] + O(N^{-7/6+2k}). \tag{5.10}
\]
Here, we have replaced the factor $\|g_a\|^{-2} \ell^2$ by 1 using $\ell^2 \|g_a\|^{-2} = 1 + O(N^{-1/2})$ and the matrix $R_a$ by the identity using
\[
\text{tr}(\tilde{B}G\tilde{B}^k) \sim r_0^a \tilde{B}G\tilde{B}^k r_a = O(N^{-1+c}), \quad k = 0, 1. \tag{5.11}
\]
We remark that (5.11) can be proved using the Ward identity and (5.10).

For the second term of (5.9), we first consider the case where $c = a$;
\[
\frac{1}{N} \sum_{c \neq a} \mathbb{E}[F'(Y) \frac{\ell^2}{\|g_a\|^2} e^{c} \tilde{B}^{(a)} R_a G[e_c(e_a + h_a)^*, \tilde{B}]G_{ea}] = \frac{1}{N} r_0^a \tilde{B}G_{ea} = \frac{1}{N} b_a h_a^* G_{ea} e_a(e_a + h_a)^*, \tilde{B}]G_{ea} = O(N^{-5/3+2k}).
\]
Then we use
\[
\text{tr}(\tilde{B}G_{eb}e^* \tilde{G}^k) \sim r_0^a \tilde{B}G_{eb}e^* \tilde{G}^k r_a = O(N^{-5/3+2k}), \quad k = 1, 2,
\]
to obtain
\[
\gamma \frac{1}{N} \sum_{c \neq a} \mathbb{E}[F'(Y) \frac{\ell^2}{\|g_a\|^2} e^{c} \tilde{B}^{(a)} R_a G[e_c(e_a + h_a)^*, \tilde{B}]G_{ea}] \\
= \gamma \mathbb{E}[F'(Y) \frac{(\tilde{B}G)bb}{N} \left((\tilde{B}G)_{ab} + h_a^* G_{eb}G_{ba}\right)] - \gamma \mathbb{E}[F'(Y) \frac{(\tilde{B}G^2)bb}{N} (G_{ab} + h_a^* G_{eb})] + O(N^{-7/6+2k}).
\]
We find from Lemma 4.10 that the remaining terms in (5.9) are negligible since all of them have additional small factors such as $\tilde{h}_{abc}$ or $h_{abc}$. For simplicity, we omit further details for these terms. Combining above results and recalling (5.2) and (5.3), we can write
\[
\mathbb{E}[F'(Y)\tilde{B}G_{ab}G_{ba}] = (1 + \tilde{w}_b \tilde{m}) \mathbb{E}[F'(Y)\tilde{B}G_{ab}G_{ba}] \\
+ (b_a - \tilde{w}_b)(1 + \tilde{w}_b \tilde{m}) \mathbb{E}[F'(Y)h_a^* G_{ebG_{ba}}] - \tilde{w}_b(1 + \tilde{w}_b \tilde{m}) \mathbb{E}[F'(Y)G_{ebG_{ba}}] \\
- \mathbb{E}[F'(Y)\tilde{B}G_{ab}G_{ba}] + \mathbb{E}[F'(Y)\tilde{B}G_{ab}G_{ba}] + \mathbb{E}[F'(Y)(b_a \tilde{B} - \tilde{B})h_a^* G_{ebG_{ba}}] \\
+ \gamma \mathbb{E}[F'(Y) \frac{(\tilde{B}G^2)bb}{N} \tilde{G}_{aa}] + \gamma \mathbb{E}[b_a F'(Y) \frac{(\tilde{B}G^2)bb}{N} h_a^* G_{ea}] \tag{5.12}
\]
\[
- \gamma \mathbb{E}[F'(Y) \frac{(\tilde{B}G^2)bb}{N} G_{aa}] - \gamma \mathbb{E}[F'(Y) \frac{(\tilde{B}G^2)bb}{N} h_a^* G_{ea}] \\
- \frac{1}{N} \sum_{c \neq a} \mathbb{E}[F'(Y) \frac{\partial}{\partial y_{ac}} \tilde{B}G_{cb}G_{ba}] + O(N^{-7/6+2k}).
\]

Step 2: Expansion of $\mathbb{E}[F'(Y)h_a^* G_{ebG_{ba}}]$  

Now we expand $\mathbb{E}[F'(Y)h_a^* G_{ebG_{ba}}]$ using the same method. Specifically, we write
\[
\mathbb{E}[F'(Y)h_a^* G_{ebG_{ba}}] = \sum_c \mathbb{E} \left[ \frac{\partial}{\partial y_{ac}} \left( \|g_a\|^{-1} F'(Y)G_{ebG_{ba}} \right) \right] + O(N^{-7/6+2k}).
\]
Now the same argument as in the proof of (5.12) leads to
\[
E[F'(Y)h_a^*G_bG_{ba}] = -\gamma E[F'(Y) \left( \text{tr}(BG)_{ab} + b_a h_a^* G_{ba} \right) G_{ba}]
\]
\[
- \gamma E[F'(Y) \left( \frac{(G^2)_{bb}}{N}((BG)_{aa} + b_a h_a^* G_{a}) - \frac{(BG)_{bb}}{N}(G_{aa} + h_a^* G_{a}) \right)]]
\]
\[
+ \frac{1}{N} \sum_c E[F''(Y)G_{cb}G_{ba} \frac{\partial Y}{\partial g_{ac}}] + O(N^{-7/6+2\epsilon}).
\]
As before, we extract the deterministic factors from tracial quantities to get
\[
E[F'(Y)h_a^*G_bG_{ba}] = -\bar{m}B_{a,b}^{(2)} + (1 + \bar{\omega}_a \bar{m}) \bar{B}^{(1)}_{a} + (1 + \bar{\omega}_a \bar{m} - \bar{m} b_a)E[F'(Y)h_a^*G_bG_{ba}]
\]
\[
- \gamma E[F'(Y) \left( \frac{(G^2)_{bb}}{N}((BG)_{aa} + b_a h_a^* G_{a}) - \frac{(BG)_{bb}}{N}(G_{aa} + h_a^* G_{a}) \right)]]
\]
\[
+ \frac{1}{N} \sum_c E[F''(Y)G_{cb}G_{ba} \frac{\partial Y}{\partial g_{ac}}] + O(N^{-7/6+2\epsilon}).
\]
Using (5.13) to substitute \(E[F'(Y)h_a^*G_bG_{ba}]\) in (5.12) concludes the proof of (5.7). \(\square\)

Next, we express (5.7) as deterministic linear combinations of \(x_h\) and \(Z_{kl}\) using the two lemmas below.

**Lemma 5.3.** Under the conditions in Proposition 5.7 the following holds true uniformly over \(a \in [1, N]\) and \(k = 1, 2, 3:\n
\[
E[F'(Y)(BG)_{aa} \text{ Tr}(BG^{k-1}G)] = \frac{(\bar{\omega}_a + \bar{m})}{N(\bar{\omega}_a - \bar{\omega}_a)} x_h
\]
\[
+ \frac{(\bar{\omega}_a + \bar{m})}{\gamma(\bar{\omega}_a - \bar{\omega}_a)^2} - \frac{1}{\gamma(\bar{\omega}_a - \bar{\omega}_a)} \right) v_a^* Z e_k + O(N^{5/6+2\epsilon}),
\]
\[
E[F'(Y)h_a^*G_{ba} \text{ Tr}(BG^{k-1}G)] = O(N^{5/6+2\epsilon}),
\]
\[
E[F'(Y)G_{ba} \text{ Tr}(BG^{k-1}G)] = \frac{1}{\gamma(\bar{\omega}_a - \bar{\omega}_a)} x_h + \frac{1}{\gamma(\bar{\omega}_a - \bar{\omega}_a)} v_a^* Z e_k + O(N^{5/6+2\epsilon}).
\]

**Lemma 5.4.** Under the conditions in Proposition 5.1 the following holds true uniformly over \(a \in [1, N]\) and \(k = 1, 2, 3:\n
\[
E[F'(Y)\partial h_a(\bar{B}G^2)_{aa}] = \frac{\bar{\omega}_a + \bar{m}}{N(\bar{\omega}_a - \bar{\omega}_a)} e_{a}^* Z v_{\beta} - \frac{1}{N(\bar{\omega}_a - \bar{\omega}_a)} e_{a}^* Z v_{\beta} + O(N^{-1/6+2\epsilon}),
\]
\[
E[F'(Y)\partial h_a^* G^2 e_a] = \frac{e_{a}^* Z v_{\beta}}{N(\bar{\omega}_a - \bar{\omega}_a)(\bar{\omega}_a - \bar{\omega}_a)} + \frac{\bar{m}}{N(\bar{\omega}_a - \bar{\omega}_a)(\bar{\omega}_a - \bar{\omega}_a)}(\bar{\omega}_a + \bar{m})Z_{k1} + Z_{k2}
\]
\[
+ O(N^{-1/6+2\epsilon}).
\]

Applying Lemmas 5.3 and 5.4 to (5.7), we now have finished decoupling the index \(a\) from \(E[F'(Y)(\bar{B}G^2)_{aa}]\).

Next, we express the quantities \(v_a^* Z e_k\) and \(e_{a}^* Z v_{\beta}\) in terms of \(Z_{11}\). The following lemma enables such conversion:

**Lemma 5.5.** Under the conditions in Proposition 5.1 the following hold true for each \(i, j = 1, 2, 3: \n
\[
Z_{ij} = \frac{u_{i,j}}{u_{j,j}} Z_{11} + O(N^{5/6+2\epsilon}) = \frac{u_{i,j}}{u_{j,j}} Z_{11} + O(N^{5/6+2\epsilon})
\]

Applying Lemma 5.5 to functional of \(Z\) that appear in Lemmas 5.3 and 5.4 yields
\[
v_a^* Z e_k = v_a^* Z e_1 \frac{u_{a,b}}{u_{b,b}} + O(N^{5/6+2\epsilon}), \quad e_{a}^* Z v_{\beta} = e_{a}^* Z v_{\beta} + O(N^{5/6+2\epsilon}),
\]
\[
K_{kl} = K_{kl} + O(N^{5/6+2\epsilon}).
\]

Finally, we deal with the last term of (5.7) in the following lemma.
Lemma 5.6. Under the conditions in Proposition 5.1, for \( k = 1, 2 \), we have the following:

\[
\sum_c \left( a \right) E[F''(Y) \frac{\partial Y}{\partial y_{ac}} (\overline{B}^{k-1}G^2)_{ac}]
= \gamma \sum_{\beta_1} \frac{1}{u_{\beta_1}} \left( \frac{\omega_{\beta_1} - m - 1}{N(a_{\beta_1} - \omega_{\alpha})} \right)^2 \frac{u_{\beta_1} u_{\beta_2}}{u_{\beta_1}^2} \int_{E_1} E[F''(Y) \text{Tr} G^2 \text{Im}(G^2)] dE + O(N^{2/3+2\epsilon}).
\]

We postpone the proofs of Lemmas 5.3, 5.6 to Section 9. We conclude this section with the proof of Proposition 5.1.

Proof of Proposition 5.1. In order to conclude the proof, we apply above lemmas to each term of (5.7).

By (5.17) and (5.20), the second term of (5.7) can be written as

\[
\text{On the other hand, reduces to}
\]

The fifth term of (5.7) can be represented as

\[
\text{The sixth term of (5.7) can be written as}
\]

The seventh term of (5.7) is \( O(N^{-1/6+2\epsilon}) \) by Lemma 5.3. Finally, by Lemma 5.6, the last term of (5.7) reduces to

\[
\frac{1}{N(a_{\beta_1} - \omega_{\alpha})^2} \left( (\tilde{\omega}_{\beta_1} + \tilde{m} - 1) x_2 - x_3 \right) + \frac{1}{N(a_{\beta_1} - \omega_{\alpha})^2} \left( (\tilde{\omega}_{\beta_2} + \tilde{m} - 1) u_{\beta_2} - u_{\beta_3} \right) \nu_{\beta_1} Z e_1 + O(N^{-1/6+2\epsilon}).
\]

Finally, we have

\[
E[F'(Y) ((\nu_{\beta} - \tilde{\omega}_{\beta}) d_1 - (\nu_{\beta} - \tilde{\omega}_{\beta})(\tilde{m} - 1) d_1 - (d) h_{\nu}^2 G^2 e_1) = O(N^{-1/6+2\epsilon}).
\]

We can write the fourth term of (5.7) as

\[
\frac{1}{N(a_{\beta_1} - \omega_{\alpha})} \left( (\tilde{\omega}_{\beta_1} + \tilde{m} - 1) x_2 - x_3 \right) = \frac{1}{N(a_{\beta_1} - \omega_{\alpha})} \left( (\tilde{\omega}_{\beta_2} + \tilde{m} - 1) u_{\beta_2} - u_{\beta_3} \right) \nu_{\beta_1} Z e_1 + O(N^{-1/6+2\epsilon}).
\]

On the other hand,

\[
v_{\beta}^2 u_{\beta} - \tilde{m} - 1(\tilde{\omega}_{\beta} + \tilde{m} - 1) u_{\beta_1} + \tilde{m} - 1 u_{\beta_2} = \frac{1}{N} \sum_a \left( \frac{b_a - \tilde{\omega}_{\beta} - \tilde{m} - 1}{b_a - \omega_{\beta}} \right) = 0.
\]

Hence we have

\[
E[F'(Y) ((\nu_{\beta} - \tilde{\omega}_{\beta}) d_1 - (\nu_{\beta} - \tilde{\omega}_{\beta})(\tilde{m} - 1) d_1 - (d) h_{\nu}^2 G^2 e_1) = O(N^{-1/6+2\epsilon}).
\]
Collecting all these results, we obtain
\[
E[F'(Y)(\tilde{B}G^2)_{a_1}] = (\tilde{m}^{-1} + \tilde{\omega})x_{a_1} - \left(\frac{1}{N} - \tilde{\omega}\right) v_{\beta}^T x
- \frac{v_{\beta}^T u_{\beta}}{N u_{\beta}(a_\beta - \tilde{\omega})^2} v_{\beta}^T Z e_{1} + 2 \frac{v_{\beta}^T u_{\beta}}{N u_{\beta}(a_\beta - \tilde{\omega})^2} \left(\tilde{\omega} + \tilde{m}^{-1}\right) Z_{11} - Z_{21}
+ \frac{v_{\beta}^T u_{\beta}}{N u_{\beta}(a_\beta - \tilde{\omega})^2} \left(\int_{E_1} E[F'(Y) Tr G^2 Im(G^2)]dE\right) \left(- \frac{1}{N(a_\beta - \tilde{\omega})^2} + \frac{\tilde{m}}{N(a_\beta - \tilde{\omega})}\right) (\tilde{m}^{-2} u_{\beta 1} - 1)
+ O(N^{-7/6+2\epsilon}).
\]  

(5.21)

Finally, Lemma 5.5 gives
\[
\tilde{m}^{-1}(\tilde{\omega} + \tilde{m}^{-1})Z_{11} = v_{\beta}^T Z e_{1} = v_{\beta}^T u_{\beta} Z_{11}.
\]  

(5.22)

Substituting (5.22) into (5.21) and using the fact that \(v_{\beta}^T u_{\beta} = \tilde{m}^{-2} u_{\beta 1} - 1\), we conclude the proof of Proposition 5.1. \(\square\)

6 Decoupling lemmas for remainders

In this section, we prove Lemmas 5.3, 5.5. Proofs in this section are parallel to that of Lemma 5.2 in the sense that we expand two different quantities, one involving \((\tilde{B}G)_{ab}\) and the other concerning \(h_a^* G e_b\).

6.1 Proof of Lemma 5.4

As mentioned at the beginning of this section, we apply Stein’s lemma to two different quantities:
\[
E[F'(Y)\partial_1(\tilde{B}G)_{a_1} G_{b_1}], \quad E[F'(Y)\partial_1 h_a^* G e_b G_{b_1}].
\]  

(6.1)

In order to expand the first quantity in (6.1), we write
\[
E[F'(Y)\partial_1(\tilde{B}G)_{a_1} G_{b_1}] = -\frac{1}{N} \sum_c E[F'(Y)\|g_c\|^{-1} \left(\frac{\partial}{\partial g_{ac}} \tilde{B}^{(a)} R_a G e_b G_{b_1}\right)
- \frac{1}{N} \sum_c E[F'(Y)\|g_c\|^{-1} \partial_1 e_c^* \tilde{B}^{(a)} R_a e_b G_{b_1}\]
- \frac{1}{N} \sum_c E[F'(Y)\|g_c\|^{-1} \partial_1 e_c^* \tilde{B}^{(a)} R_a G e_b e_b^* G_{b_1}\]
- \frac{1}{N} \sum_c E[F'(Y)\|g_c\|^{-1} \partial_1 e_c^* \tilde{B}^{(a)} R_a G e_b e_b^* G_{b_1}\] + O(N^{-7/6+2\epsilon}).
\]  

(6.2)

To calculate the derivative of \(\partial_1\), note that \(\text{tr}[X,Y]Z = \text{tr}[X,Y]\) implies
\[
\text{tr} \frac{\partial \tilde{B}^{k-1}}{\partial g_{ac}} G = \text{tr} \left[\frac{\partial}{\partial g_{ac}} (R_a U^{(a)} \tilde{B}^{k-1} (U^{(a)})^* R_a) G\right] = \text{tr} \left[\frac{\partial}{\partial g_{ac}} R_a, \tilde{B}^{k-1}\right] G
= \frac{\ell^2}{\|g_c\|} \text{tr} \left(e_c (e_a + h_a) \ast - \frac{h_a e_a}{2} (e_a + 2h_a) e_a + \frac{h_a e_a}{2} (e_a e_a + h_a e_a + e_a h_a)\right) \tilde{B}^{k-1}, G
= O(N^{-4/3+2\epsilon}).
\]  

(6.3)

and
\[
\text{tr} \tilde{B}^{k-1} \frac{\partial G}{\partial g_{ac}} = -\frac{1}{N} \left(e_a + h_a\right)^* \tilde{B} G \tilde{B}^{k-1} G e_c + O(N^{-7/6+2\epsilon}) = O(N^{-2/3+2\epsilon}).
\]  

(6.4)

In (6.1), we used the fact that terms of the form \(u^* [\tilde{B}, G \tilde{B}^{k-1} G] v\) are \(O(1/3)\), where \(u\) and \(v\) can be \(e_a, h_a,\) or \(e_c\). By (6.3) and (6.4), we have that \(\frac{\partial}{\partial g_{ac}} = O(N^{-2/3+2\epsilon}).\)

Applying Lemma 4.4 to derivatives of \(G\) in (6.1), we have
\[
E[F'(Y)\partial_1 e_a^* \tilde{B} G e_b G_{b_1}] = \gamma E[F'(Y)\partial_1 \text{tr}(\tilde{B}G) e_a^* \tilde{B} G e_b G_{b_1}] + \gamma E[b_a F'(Y)\partial_1 \text{tr}(\tilde{B}G) h_a^* G e_b G_{b_1}]
- \gamma E[F'(Y)\partial_1 \text{tr}(\tilde{B}G) e_a^* Ge_b G_{b_1}] - \gamma E[F'(Y)\partial_1 \text{tr}(\tilde{B}G) h_a^* Ge_b G_{b_1}]
+ \gamma E[F'(Y)\partial_1 \left(\frac{\langle GB \rangle_{(G)}}{N} e_a^* \tilde{B} G e_a\right) + \gamma E[b_a F'(Y)\partial_1 \left(\frac{\langle GB \rangle_{(G)}}{N} h_a^* G e_a\right]
- \gamma E[F'(Y)\partial_1 \left(\frac{\langle G \tilde{B}^2 \rangle_{(G)}}{N} e_a^* \tilde{B} G e_a\right) - \gamma E[F'(Y)\partial_1 \left(\frac{\langle G \tilde{B}^2 \rangle_{(G)}}{N} h_a^* G e_a\right) + O(N^{-7/6+2\epsilon}),
\]  

(6.5)
so that
\[
E[F'(Y)\partial_k e^*_a \tilde{B} G e_b G_{ba}] = (1 + \bar{\omega}_\beta \bar{m})(1 + \bar{\omega}_\beta \bar{m})E[F'(Y)\partial_k (\tilde{B} G)_{ab} G_{ba}]
\]
\[
+ (1 + \bar{\omega}_\beta \bar{m})(1 + \bar{\omega}_\beta \bar{m})E[F'(Y)\partial_k h^*_a G e_b G_{ba}] - \bar{\omega}_\beta(1 + \bar{\omega}_\beta \bar{m})E[F'(Y)\partial_k G_{ab} G_{ba}]
\]
\[
+ \frac{\bar{\omega}_\beta + \bar{m}^{-1}}{N(a_a - \bar{\omega}_a)} \sum_{kb2} - \frac{1}{N(a_a - \bar{\omega}_a)} Z_{kh2} + O(N^{-7/6 + 2\epsilon}).
\]

The second quantity of (6.3) can be expanded in the same way, which leads to
\[
E[F'(Y)\partial_k h^*_a G e_b G_{ba}] = -\bar{m}E[F'(Y)\partial_k (\tilde{B} G)_{ab} G_{ba}] - \bar{m}E[F'(Y)\partial_k h^*_a G e_b G_{ba}]
\]
\[
+ (1 + \bar{\omega}_\beta \bar{m})E[F'(Y)\partial_k G_{ab} G_{ba}] + (1 + \bar{\omega}_\beta \bar{m})E[F'(Y)\partial_k h^*_a G e_b G_{ba}]
\]
\[
- \gamma E[F'(Y)\partial_k (\tilde{B} G)_{ab} G_{ba}] - \gamma E[F'(Y)\partial_k h^*_a G e_b G_{ba}]
\]
\[
+ \gamma E[F'(Y)\partial_k (\tilde{B} G)_{ab} G_{ba}] + \gamma E[F'(Y)\partial_k (\tilde{B} G)_{ab} G_{ba}] + O(N^{-7/6 + 2\epsilon}),
\]
so that
\[
(b_a - \bar{\omega}_\beta)E[F'(Y)\partial_k h^*_a G e_b G_{ba}] = -\bar{m}E[F'(Y)\partial_k (\tilde{B} G)_{ab} G_{ba}] + (\bar{\omega}_\beta + \bar{m}^{-1})E[F'(Y)\partial_k G_{ab} G_{ba}]
\]
\[
E[F'(Y)\partial_k (\tilde{B} G)_{ab} G_{ba}] = \bar{m}^{-1}(\bar{\omega}_\beta + \bar{m}^{-1})Z_{kh1} + \frac{\bar{m}^{-1}}{N(a_a - \bar{\omega}_a)} Z_{kh2} + O(N^{-7/6 + 2\epsilon}).
\]

Combining the equations (6.4) and (6.6), we obtain that
\[
E[F'(Y)\partial_k (\tilde{B} G)_{ab} G_{ba}] = (\bar{\omega}_\beta + \bar{m}^{-1})E[F'(Y)\partial_k G_{ab} G_{ba}] - \frac{1}{N(a_a - \bar{\omega}_a)} e^*_a Z_{v} + O(N^{-7/6 + 2\epsilon}),
\]
and that
\[
E[F'(Y)\partial_k h^*_a G e_b G_{ba}] = \frac{1}{N(a_a - \bar{\omega}_a)}(b_a - \bar{\omega}_\beta) e^*_a Z_{v}.
\]

Taking the sum of (6.5) over b proves (6.18).

Note that the first term of (6.7) still has the index a within the expectation. In the rest of the proof, we derive the decoupled form of \(\partial_k G_{ab} G_{ba}\) using a similar argument. We write
\[
z E[F'(Y)\partial_k G_{ab} G_{ba}] = \gamma \bar{m}E[F'(Y)\partial_k G_{ab} G_{ba}] + \gamma E[F'(Y)\partial_k (\tilde{B} G)_{ab} G_{ba}]
\]
\[
+ \gamma \sqrt{t}E[F'(Y)\partial_k (W G)_{ab} G_{ba}] - \delta_{a_b} E[F'(Y)\partial_k G_{ab} G_{ba}] + \gamma \sqrt{t}E[F'(Y)\partial_k (W G)_{ab} G_{ba}] - \delta_{a_b} E[F'(Y)\partial_k G_{ab} G_{ba}].
\]

Note that the third term can be written as
\[
\gamma \sqrt{t} \sum_c E[F'(Y)\partial_k W_{ac} G_{cb} G_{ba}] = \frac{\gamma \sqrt{t}}{\sqrt{N}} \sum_c E[F'(Y)\partial_k G_{ab} G_{ba}]
\]
\[
= \frac{\gamma \sqrt{t}}{\sqrt{N}} \sum_c E[F'(Y)\partial_k G_{ab} G_{ba} + F'(Y)\partial_k G_{cb} G_{ba} + F'(Y)\partial_k G_{cb} G_{ba} + F'(Y)\partial_k G_{cb} G_{ba}]
\]
\[
= O(N^{-4/3 + 2\epsilon}).
\]

Substituting (6.7) into the second term of (6.9) and rearranging the equation, we have
\[
E[F'(Y)\partial_k G_{ab} G_{ba}] = \delta_{a_b} \frac{1}{N(a_a - \bar{\omega}_a)^2} E[F'(Y)\partial_k G_{ab} G_{ba}] + \frac{1}{N(a_a - \bar{\omega}_a)^2} e^*_a Z_{v} + O(N^{-7/6 + 2\epsilon}),
\]
where we have used the identity
\[
\frac{z}{\gamma} - \bar{m}^{-1} - \bar{\omega}_\beta - a_a = \bar{\omega}_\beta - a_a - t \bar{m}
\]
from the subordination equation (5.2). Plugging (6.10) into (6.7) and taking the sum over b yields
\[
E[F'(Y)\partial_k (\tilde{B} G)^a_{ab} G_{ba}] = \frac{\bar{\omega}_\beta + \bar{m}^{-1}}{N(a_a - \bar{\omega}_a)^2} e^*_a Z_{v} + \frac{1}{N(a_a - \bar{\omega}_a)^2} e^*_a Z_{v} + O(N^{-1/6 + 2\epsilon}),
\]
which proves (5.17).
6.2 Proof of Lemma [5.3]

As in the previous section, we expand the following two quantities for $k = 0, 1, 2$;

$$
\mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb}(\tilde{B}G)_{aa}], \quad \mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb}h_a G_e a].
$$

(6.11)

We only demonstrate the terms of leading order, and the remainders can be bounded following the same lines as in the previous section. To expand the first term of (6.11), we write

$$
\begin{align*}
\mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb}(\tilde{B}G)_{aa}] &= \frac{\gamma}{N} \sum_c \mathbb{E}[F'(Y)\|g_a\|^2 - \hat{\ell}_c^2 \langle (G\tilde{B}^k G)_{bb} G_c (e_c + h_a)^* \rangle, \hat{B} \hat{G} e_a] + O(N^{-7/6+\epsilon}) \\
&= -\gamma \mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb} tr(\tilde{B}G)(\tilde{B}G)_{aa}] + (1 + \tilde{\omega}_{\beta}\tilde{m})(b_a - \tilde{\omega}_{\beta}) \mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb} h^*_a G_e a] \\
&\quad - \gamma \mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb} tr(\tilde{B}G)\hat{G} G_a] + O(N^{-7/6+\epsilon}) \\
&= (1 + \tilde{\omega}_{\beta}\tilde{m})\mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb} (\tilde{B}G)_{aa}] + (1 + \tilde{\omega}_{\beta}\tilde{m})(b_a - \tilde{\omega}_{\beta}) \mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb} h^*_a G_e a] \\
&\quad - \tilde{\omega}_{\beta}(1 + \tilde{\omega}_{\beta}\tilde{m})\mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb} G_a] + \frac{\tilde{\omega}_{\beta} + \tilde{m}^{-1}}{\gamma(\alpha_a - \omega_a)} \mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb} \hat{G} e_x] + O(N^{-1/6+2\epsilon}).
\end{align*}
$$

(6.12)

On the other hand, the second term of (6.11) can be written as follows:

$$
\begin{align*}
\mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb} h^*_a G_e a] &= -\frac{\gamma}{N} \sum_c \mathbb{E}[F'(Y)\|g_a\|^2 - \hat{\ell}_c^2 \langle (G\tilde{B}^k G)_{bb} G_c (e_c + h_a)^* \rangle, \hat{B} \hat{G} e_a] \\
&= -\gamma \mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb} tr(\tilde{B}G)(\tilde{B}G)_{aa}] - \tilde{m}\mathbb{E}[b_a F'(Y)(G\tilde{B}^k G)_{bb} h^*_a G_e a] \\
&\quad + \gamma \mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb} tr(\tilde{B}G)\hat{G} G_a] + (1 + \tilde{\omega}_{\beta}\tilde{m})\mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb} h^*_a G_e a] + O(N^{-1/6+2\epsilon}).
\end{align*}
$$

(6.13)

Rearranging this equation yields

$$
\begin{align*}
(b_a - \tilde{\omega}_{\beta})\mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb} h^*_a G_e a] &= -\gamma \mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb} tr(\tilde{B}G)(\tilde{B}G)_{aa}] + (\tilde{\omega}_{\beta} + \tilde{m}^{-1})\mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb} h^*_a G_e a] \\
&\quad - \frac{\tilde{m}^{-1}(\tilde{\omega}_{\beta} + \tilde{m}^{-1})}{\gamma(\alpha_a - \omega_a)} \mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb} \hat{G} e_x] + O(N^{-1/6+2\epsilon}).
\end{align*}
$$

(6.14)

Substituting (6.13) into (6.12) we obtain

$$
\begin{align*}
\mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb} (\tilde{B}G)_{aa}] &= (\tilde{\omega}_{\beta} + \tilde{m}^{-1})\mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb} G_a] - \frac{\sqrt{\eta} Z e_k}{\gamma(\alpha_a - \omega_a)} + O(N^{-1/6+2\epsilon}), \quad (6.15)
\end{align*}
$$

and plugging in this back to (6.13) gives

$$
\begin{align*}
\mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb} h^*_a G_e a] &= \frac{1}{\gamma(\alpha_a - \omega_a)} \mathbb{E}[F'(Y)\hat{\theta}(G\tilde{B}^k G)_{bb}] + O(N^{-1/6+2\epsilon}).
\end{align*}
$$

(6.16)

Note that (6.15) together with $\langle \theta \rangle < N^{-1/2+\epsilon}$ proves (6.15).

As in the previous section, we still need to decouple the index $a$ from the first term on the right-hand side of (6.14). To this end, we write

$$
\begin{align*}
\gamma \mathbb{E}[F'(Y)G_{aa}(G\tilde{B}^k G)_{bb}] &= \gamma \mathbb{E}[F'(Y)G_{aa}(G\tilde{B}^k G)_{bb}] + \gamma \mathbb{E}[F'(Y)(\tilde{B}G)_{aa}(G\tilde{B}^k G)_{bb}] \\
&\quad + \gamma \sqrt{\eta} \mathbb{E}[F'(Y)(W G)_{aa}(G\tilde{B}^k G)_{bb}] - \mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb}] + O(N^{-1/6+2\epsilon}).
\end{align*}
$$

(6.17)

Substituting (6.14) into the second term of (6.16) and rearranging the equation, we have

$$
\begin{align*}
\mathbb{E}[F'(Y)G_{aa}(G\tilde{B}^k G)_{bb}] &= \frac{1}{\gamma(\alpha_a - \omega_a)} \mathbb{E}[F'(Y)(G\tilde{B}^k G)_{bb}] + \frac{1}{\gamma(\alpha_a - \omega_a)^2} \sqrt{\eta} Z e_k + O(N^{-1/6+2\epsilon}), \quad (6.17)
\end{align*}
$$

(6.18)
where we have used
\[
\sqrt{T} \sum_c \mathbb{E}[F'(Y)W_{ac}G_{ca}(GB\tilde{b}^{\delta}G)_{ab}] = \gamma \frac{1}{N} \sum_c \mathbb{E}[\frac{\partial}{\partial X_{ca}}(F'(Y)G_{ca}(GB\tilde{b}^{\delta}G)_{ab})] \\
= -tm\mathbb{E}[F'(Y)G_{ao}(GB\tilde{b}^{\delta}G)_{ab}] + O(N^{-2/3+2\epsilon}).
\]

Taking the sum of (6.17) over \(b\) proves (5.16), and plugging (6.17) into (6.14) gives (5.14). This concludes the proof of Lemma 5.5.

### 6.3 Proof of Lemma 5.5

In this section, we prove Lemma 5.5 that allows us to express the matrix \(Z\) and \(\tilde{Z}\) in terms of the vector \(Ze_1 = \tilde{Z}e_1\). We first claim that the following analogue of (6.10) holds true:
\[
\mathbb{E}[F'(Y)\partial_k G_{ba}] = \delta_{ab} \frac{1}{\gamma(b_a - \omega_{a})} \mathbb{E}[F'(Y)\partial_k G_{a}] + \frac{1}{N(b_a - \omega_{a})^2} \varepsilon_k^T \tilde{Z}_a \varepsilon_a + O(N^{-7/6+2\epsilon}). \tag{6.18}
\]

Interchanging the roles of \(A\) and \(B\), it suffices to prove
\[
\mathbb{E}[F'(Y)d_k G_{ba}] = \delta_{ab} \frac{1}{\gamma(a_b - \omega_{a})} \mathbb{E}[F'(Y)d_k G_{a}] + \frac{1}{N(a_b - \omega_{a})^2} \varepsilon_k \tilde{Z}_b \varepsilon_b + O(N^{-7/6+2\epsilon}). \tag{6.19}
\]

We can easily see that the proof of (6.10) applies almost verbatim to (6.19), with the only difference being (6.10) replaced by
\[\operatorname{tr} \frac{\partial A^k}{\partial \eta_{ac}} G = 0.\]

We omit the details to avoid repetition.

Taking the sum of (6.18) over \(a, b\) with weights \(a_{i-1}^i\), we have
\[
Z_{ki} = \mathbb{E}[F'(Y)\partial_k \operatorname{Tr} B^{i-1}G^2] = \mathbb{E}[F'(Y)\partial_k \operatorname{Tr} B^{i-1}G^2] = u_{\beta i} e_k^i \tilde{Z}_\nu \varepsilon_a + O(N^{5/6+2\epsilon}). \tag{6.20}
\]

Then we obtain from (6.20) that
\[
u_{\beta j} Z_{kj} - u_{\beta j} Z_{k1} = e_k^i Z(u_{\beta i} e_j - u_{\beta j} e_i)^T = O(N^{5/6+2\epsilon}), \quad j = 1, 2, 3. \tag{6.21}
\]

This proves the first equality of (5.16).

On the other hand, taking the average over \(a\) of (5.16) with weights \(a_{i-1}^i\), we have
\[
\tilde{Z}_i = u_{\alpha i} \varepsilon_\beta Z e_1 + tm u_{\alpha i} x_1 + O(N^{5/6+2\epsilon}).
\]

By symmetry, changing the roles of \(A\) and \(B\), we have the following analogue of (6.10) for \(k = 1\):
\[
\gamma \mathbb{E}[F'(Y)G_{aa} \operatorname{Tr} G^2] = \frac{1}{(b_a - \omega_{a})^2} x_1 + \frac{1}{(b_a - \omega_{a})^2} \varepsilon_a^i Z e_1 + \frac{tm}{(b_a - \omega_{a})^2} x_1 + O(N^{5/6+2\epsilon}). \tag{6.22}
\]

Since \(\operatorname{Tr} G^2 = \operatorname{Tr} G^2, \ 3 e_1 = \tilde{3} e_1\), and \(x_1 = x_1\), taking the average over \(a\) of (6.22) with weights \(b_{a-1}^a\) gives
\[
Z_{i1} = u_{\beta i} \varepsilon_\beta Z e_1 + tm u_{\beta i} x_1 + O(N^{5/6+2\epsilon}).
\]

Therefore we have
\[
u_{\beta i} Z_{i1} - u_{\beta i} Z_{11} = O(N^{5/6+2\epsilon}), \quad i = 1, 2, 3. \tag{6.23}
\]

Combining (6.21) and (6.23), we obtain
\[
Z_{ij} = \frac{u_{\beta j} u_{\beta i}}{u_{\beta 1}^2} Z_{11} + O(N^{5/6+2\epsilon}).
\]

This concludes the proof of Lemma 5.5.
6.4 Proof of Lemma 5.6

In this section, we expand the terms in (6.2) with partial derivatives \( \frac{\partial Y}{\partial g_{ac}} \) in order to decouple the index \( a \) from them. For \( a \neq c \), by (6.1) we have

\[
\frac{\partial Y}{\partial g_{ac}} = \frac{1}{2i} \int_{E_1} \left( \frac{\partial Tr \tilde{G}}{\partial g_{ac}} - \frac{Tr \tilde{G}^*}{\partial g_{ac}} \right) dE = -\frac{\gamma}{2i} \int_{E_1} \left( e_a + h_a \right)^* \left[ \tilde{B}, \tilde{G}^2 - (\tilde{G}^*)^2 \right] e_c dE + O(N^{-5/6+2\epsilon})
\]

where \( \tilde{G} \) denotes \( G(E + L_+ + i\eta_0) \), with \( E \) being the integrating variable. Thus for \( k = 0, 1 \) we have

\[
\sum_{c} \mathbb{E}[F''(Y)e_{ca} \frac{\partial Y}{\partial g_{ac}} (\tilde{B}^kG^2)]c = \frac{\gamma}{2i} \sum_{c} \mathbb{E}[F''(Y)(\tilde{B}^kG^2)_{ca}] \int_{E_1} \left( e_a + h_a \right)^* \left[ \tilde{B}, \tilde{G}^2 - (\tilde{G}^*)^2 \right] e_c dE + O(N^{1/2+2\epsilon})
\]

\[
= -\frac{\gamma}{2i} \int_{E_1} \mathbb{E}[F''(Y)(e_a + h_a)^* \left[ \tilde{B}, \tilde{G}^2 - (\tilde{G}^*)^2 \right] \tilde{B}^kG^2 e_a] dE + O(N^{1/2+2\epsilon}). \quad (6.24)
\]

Since \( \tilde{G}^* = G(E + L_+ - i\eta_0) \), we find that it suffices to consider the following quantities with \( \tilde{G} = G(\tilde{z}) \), where \( \tilde{z} \) is a generic variable such that \( \text{Im} \tilde{z} = \pm \eta_0 \) and \( \text{Re} \tilde{z} \in [E_1, E_2] \):

\[
\mathbb{E}[F''(Y)e_{ca} \tilde{B}, \tilde{G}^2] e_a] , \quad \mathbb{E}[F''(Y)e_{ca} \tilde{B}, \tilde{G}^2] e_a] , \quad k = 0, 1.
\]

To this end, we introduce the following notation, which corresponds to \( Z \) and \( \tilde{Z} \) above; for each \( a \in [1, N] \), we define two \( (3 \times 3) \) matrices \( 3_a \) and \( \tilde{3}_a \), as

\[
3_{a_{ij}} := \mathbb{E}[F''(Y)(\tilde{B}^{-1} \tilde{G}^2 \tilde{B}^{-1} \tilde{G}^2)]_{aa}, \quad \tilde{3}_{a_{ij}} := \mathbb{E}[F''(Y)(\tilde{B}^{-1} \tilde{G}^2 \tilde{B}^{-1} \tilde{G}^2)]_{aa}, \quad i, j \in \{1, 2, 3\},
\]

where \( \tilde{G} = G(\tilde{z}) \). Similar to \( Z \) and \( \tilde{Z} \), we also define \( \tilde{3} = \sum_a 3_a \) and \( \tilde{\tilde{3}} = \sum_a \tilde{3}_a \).

Using the notations above, we now prove Lemma 5.6 from the following two lemmas, whose proofs are postpone to the end of this section:

**Lemma 6.1.** Under the conditions in Proposition 4.2, the following holds true uniformly over \( a \in [1, N] \) and \( i, j = 1, 2, 3 \):

\[
3_{a_{ij}} = \frac{1}{N(a_a - \omega_a)} e_i^a u_{\beta}^a e_j + O(N^{1/3+2\epsilon}), \quad (6.25)
\]

\[
3_{a_{ij}} = \frac{\omega_a + \tilde{m}}{N(a_a - \omega_a)} e_i^a u_{\beta}^a e_j + O(N^{1/3+2\epsilon}), \quad (6.26)
\]

\[
\mathbb{E}[F''(Y)h_{i}^{a} \tilde{G}^2 \tilde{B}^{-1} \tilde{G}^2 e_a] = \frac{(\omega_a (\tilde{m} + \tilde{m}^{-1}) e_1 - (2\tilde{m} + \tilde{m}^{-1}) e_2 + e_3)^T e_j}{N(a_a - \omega_a)(b_a - \omega_a)} + O(N^{1/3+2\epsilon}). \quad (6.27)
\]

**Lemma 6.2.** For each \( i, j \), we have the following:

\[
3_{ij} = \frac{1}{u_{\beta}^a} e_i^a u_{\beta}^a e_j + O(N^{1/3+2\epsilon}), \quad \tilde{3}_{ij} = \frac{1}{u_{\alpha}^a} e_i^a u_{\alpha}^a e_j + O(N^{1/3+2\epsilon}).
\]

Furthermore, we have

\[
\tilde{3}_{ij} = \frac{u_{\beta}^a u_{\beta}^a}{u_{\beta}^a} 3_{ij} + O(N^{1/3+2\epsilon}).
\]

Now we prove Lemma 5.6 from the lemmas above. First of all, applying Lemma 6.2 to the right-hand side of (6.27), we obtain

\[
(\omega_a (\tilde{m} + \tilde{m}^{-1}) e_1 - (2\tilde{m} + \tilde{m}^{-1}) e_2 + e_3)^T e_j = (\omega_a (\tilde{m} + \tilde{m}^{-1}) e_1 - (2\tilde{m} + \tilde{m}^{-1}) e_2 + e_3)^T u_{\beta}^a e_j + O(N^{1/3+2\epsilon})
\]

\[
= \frac{1}{N} \sum_{a} \omega_a (\tilde{m} + \tilde{m}^{-1}) - (2\tilde{m} + \tilde{m}^{-1}) b_a + b^2 Z_{ij} + O(N^{1/3+2\epsilon}) = O(N^{1/3+2\epsilon}).
\]
Thus eqn. reduces to \( O(N^{4/3+2\varepsilon}) \), so that
\[
\mathbb{E}[F''(Y)h_n^* \tilde{B} \tilde{G}^2 \tilde{B}^k G^2 e_a] = \mathbb{E}[F''(Y)h_n^* \tilde{G}^2 (b_n I - \tilde{B}) \tilde{B}^k G^2 e_a] = O(N^{4/3+2\varepsilon}). \tag{6.28}
\]

On the other hand for \( k = 1, 2 \) we have
\[
\mathbb{E}[F''(Y)e_n^* \tilde{B} \tilde{G}^2 \tilde{B}^{k-1} G^2 e_a] = \mathcal{N}_2 - \mathcal{N}_1(k+1)
= \frac{1}{N(a_n - \omega_a)^2} (\bar{\omega}_{\beta} + \bar{m}^{-1}) v_\beta^a \tilde{e}_e - \frac{1}{N(a_n - \omega_a)^2} v_\beta^a \tilde{e}_k + O(N^{4/3+2\varepsilon}) \tag{6.29}
\]
\[
= \frac{v_\beta^a u_\beta \mathcal{N}_1}{N(a_n - \omega_a)^2 u_{\beta_1}^2} ((\bar{\omega}_{\beta} + \bar{m}^{-1}) u_{\beta k} - u_{\beta(k+1)}) - \frac{v_\beta^a u_\beta \mathcal{N}_1}{N(a_n - \omega_a)^2 u_{\beta_1}^2} u_{\beta k} + O(N^{4/3+2\varepsilon}).
\]

Plugging (6.28) and (6.29) into (6.24), we have
\[
\sum_c \mathbb{E}[F''(Y) \frac{\partial Y}{\partial g_{ac}} (\tilde{B}^{k-1} G^2)_{ac}]
= \frac{v_\beta^a u_\beta}{u_{\beta_1}^2} \left( \int_{E_1}^{E_2} \mathcal{N}_1(z, w) - \mathcal{N}_1(z, \overline{\mathcal{N}}) \frac{dE}{2\pi} \right) \left( - \frac{u_{\beta k} - u_{\beta(k+1)}}{N(a_n - \omega_a)^2} + \frac{u_{\beta k}}{N(a_n - \omega_a)} \right) + O(N^{2/3+3\varepsilon}),
\]
where we denoted \( w = E + L_k + i\eta \) and \( \mathcal{N}_1 = \sum u \mathcal{N}_{ij} \). Note that for \( k = 1, 2 \) we have
\[
(\bar{\omega}_{\beta} + \bar{m}^{-1}) u_{\beta k} - u_{\beta(k+1)} = \int x^{k-1} \frac{(\bar{\omega}_{\beta} + \bar{m}^{-1} - x)}{(x - \bar{\omega}_{\beta})^2} d\mu(x) = \bar{\omega}_{\beta}^{k-1} (\bar{m}^{-1} - \bar{m}).
\]

Therefore we obtain
\[
\sum_c \mathbb{E}[F''(Y) \frac{\partial Y}{\partial g_{ac}} (\tilde{B}^{k-1} G^2)_{ac}]
= \frac{v_\beta^a u_\beta}{u_{\beta_1}^2} \left( \mathcal{N}_1 - \mathcal{N}_1 \right) \left( \frac{\bar{\omega}_{\beta}^{k-1} \bar{m} - \bar{m}^{k-1} u_{\beta_1}}{N(a_n - \omega_a)^2} + \frac{u_{\beta k}}{N(a_n - \omega_a)} \right) \int_{E_1}^{E_2} \mathbb{E}[F''(Y) \text{Tr} G^2 \text{Im}(G^2)]dE + O(N^{2/3+3\varepsilon}),
\]
concluding the proof of Lemma 5.6.

### 6.4.1 Proof of Lemma 6.1

For \( \mathcal{N}_{a2j} \), as before we need to expand two quantities;
\[
\mathbb{E}[F''(Y)e_n^* \tilde{B} \tilde{G}^2 \tilde{B}^{j-1} G^2 e_a] = \mathcal{N}_{a2j}, \quad \mathbb{E}[F''(Y)h_n^* \tilde{G}^2 \tilde{B}^{j-1} G^2 e_a]. \tag{6.30}
\]

Expanding the first quantity of (6.30) yields
\[
- \sum_b \mathbb{E}[F''(Y) \frac{\bar{\omega}_{\beta k}}{\|g_a\|^2} e_b^* \tilde{B} (b_n) R_a \tilde{G}^2 \tilde{B}^{j-1} G^2 e_a]
= \gamma \sum_b \mathbb{E}[F''(Y) \frac{e_b^* \tilde{B} (b_n)}{\|g_a\|^2} R_a \tilde{G} e_b (e_a + h_a)^* \tilde{B} \tilde{G}^2 \tilde{B}^{j-1} G^2 e_a]
+ \gamma \sum_b \mathbb{E}[F''(Y) \frac{e_b^* \tilde{B} (b_n)}{\|g_a\|^2} R_a \tilde{G} e_b (e_a + h_a)^* \tilde{B} \text{Tr} G^2 \text{Im}(G^2) e_a] + O(N^{4/3+2\varepsilon}) \tag{6.31}
\]
\[
= (\bar{\omega}_{\beta} + \bar{m}^{-1}) \mathcal{N}_{a2j} - \bar{\omega}_{\beta} \mathcal{N}_{a2j} + \bar{\omega}_{\beta} \tilde{m} \mathcal{N}_{a2j} + \frac{\bar{\omega}_{\beta} + \bar{m}^{-1}}{N(a_n - \omega_a)^2} \mathcal{N}_{a2j} - \frac{1}{N(a_n - \omega_a)^2} \mathcal{N}_{a2j} + \frac{\bar{\omega}_{\beta} + \bar{m}^{-1}}{N(a_n - \omega_a)^2} \mathcal{N}_{a2j} + \frac{(\bar{\omega}_{\beta} + \bar{m}^{-1}) (\bar{\omega}_{\beta} + \bar{m}^{-1})}{N(a_n - \omega_a)^2} \mathcal{N}_{a2j} + O(N^{4/3+2\varepsilon}).
\]

Similarly, the second quantity (6.30) can be expanded as
\[
\sum_b \mathbb{E}[F''(Y) \frac{\bar{\omega}_{\beta k}}{\|g_a\|^2} e_b^* \tilde{G}^2 \tilde{B}^{j-1} G^2 e_a] = \mathcal{N}_{a2j} - \mathcal{N}_{a2j} + \frac{1}{N(a_n - \omega_a)} \left( (\bar{\omega}_{\beta} + \bar{m}^{-1}) \mathcal{N}_{a2j} - \mathcal{N}_{a2j} \right) + O(N^{4/3+2\varepsilon}). \tag{6.32}
\]
Rearranging (6.32) gives
\[
\tilde{m}(b_a - \tilde{\omega}_a)\mathbb{E}[F''(Y)h_aG^2\tilde{B}^{j-1}G^2e_a] = ((\tilde{\omega}_a + \tilde{m}^{-1})e_1 - e_2)^\top \left(\tilde{m}\tilde{3}_a - \frac{1}{N(a_a - \tilde{\omega}_a)^2}3_a e_j + O(N^{4/3+2\epsilon})\right), \tag{6.33}
\]
and plugging (6.33) into (6.31) yields
\[
3_{a_2j} = (\tilde{\omega}_2 + \tilde{m}^{-1})3_{a_1j} - \frac{1}{N(a_a - \tilde{\omega}_a)}v^\top_3 3_a e_j + O(N^{4/3+2\epsilon}). \tag{6.34}
\]
On the other hand, we also have
\[
z^\gamma(3_{a_1j}) = a_a3_{a_1j} + 3_{a_2j} + \sqrt{t}\mathbb{E}[F''(Y) e^\ast_a W G^2 \tilde{B}^{j-1}G^2e_a] - \frac{1}{\gamma}E[F''(Y) e^\ast_a \tilde{G}^2 \tilde{B}^{j-1}G^2e_a]. \tag{6.35}
\]
The third term of (6.35) is further expanded as
\[
\sqrt{t}\sum_b \mathbb{E}[F''(Y) W_{ab} e^\ast_b G^2 \tilde{B}^{j-1}G^2e_a] = -t\tilde{m}3_{a_1j} - \frac{t}{N(a_a - \tilde{\omega}_a)}\mathbb{E}[F''(Y) \text{ Tr } \tilde{G}^2 \tilde{B}^{j-1}G] + O(N^{1+\epsilon}). \tag{6.36}
\]
Substituting (6.34) and (6.36) into (6.35), leads to
\[
(\gamma^{-1} - 1 - \tilde{\omega}_a - \tilde{m})3_{a_1j} = -\frac{1}{N(a_a - \tilde{\omega}_a)}v^\top_3 3_a e_j + O(N^{4/3+2\epsilon}),
\]
which proves (6.24). We get (6.26) by combining (6.24) and (6.34), and applying (6.24) and (6.34) to (6.26) proves (6.33). This concludes the proof of Lemma 6.3.

**6.4.2 Proof of Lemma 6.2**

Taking the sum over \(a\) of (6.24) with weights \(a_a^{-1}\), we have
\[
\tilde{3}_{i,j} = e_i^\ast u_a v^\top_3 3_a e_j + O(N^{7/3+2\epsilon}).
\]
Now we claim that the following holds true:
\[
3_{i,j} = u_{i,j} v^\top_3 3_a e_j + O(N^{7/3+2\epsilon}). \tag{6.37}
\]
To prove (6.37), we first note that (6.24) remains intact for each summand if we replace \((G^2 \tilde{B}^{j-1}G^2)\) by \(\sum_{k,c} \tilde{G}^2 \tilde{B}^{j-1}G_c e_c e^\ast_c G\).

To be specific, taking \(j = 1\) we have
\[
\mathbb{E}[F''(Y) e^\ast_i G^2 e_b G_{ba} G_{ca}] = \frac{1}{N(a_a - \tilde{\omega}_a)^2} \sum_{k=1}^3 v_{ij} \mathbb{E}[F''(Y) G_b e_c G \tilde{B}^{k-1} \tilde{G}^2 e_b] + O(N^{-2/3+2\epsilon})
\]
for each \(a, b \in \{1, N\}\). By symmetry, interchanging the roles of \(A\) and \(B\) gives that
\[
\mathbb{E}[F''(Y) e^\ast_b G^2 e_b G_{ba} G_{ca}] = \frac{1}{N(a_a - \tilde{\omega}_a)^2} \sum_{k=1}^3 v_{ij} \mathbb{E}[F''(Y) G_b e_c G \tilde{B}^{k-1} \tilde{G}^2 e_b] + O(N^{-2/3+2\epsilon}). \tag{6.38}
\]
Taking the sum over \(a, b, c\) of (6.38) with weights \(b_a^{-1} b_b^{-1}\), we have
\[
3_{ij} = u_{i,j} \sum_{k=c}^3 v_{i,k} \mathbb{E}[F''(Y) \text{ Tr } \tilde{B}^{j-1}G^2 \tilde{B}^{k-1} \tilde{G}^2] = u_{i,j} v^\top_3 3_a e_j + O(N^{7/3+2\epsilon}),
\]
which is (6.37).
In summary, we have the system
\[
\begin{align*}
\tilde{3} &= u_\alpha v_\beta^\top \tilde{3} + O(N^{7/3+2\varepsilon}) \\
\tilde{3} &= u_\beta v_\alpha^\top \tilde{3} + O(N^{7/3+2\varepsilon}).
\end{align*}
\]
Solving the system gives that
\[
\tilde{3} e_j = \tilde{3} e_j = \frac{3_{1j}}{u_{j\alpha}} u_\alpha + O(N^{7/3+2\varepsilon}),
\]
\[
\tilde{3} e_j = \frac{3_{1j}}{u_{j\beta}} u_\beta + O(N^{7/3+2\varepsilon}).
\]
Furthermore, changing the roles of \(z\) and \(\tilde{z}\), we have
\[
\tilde{3} j = \frac{u_{j\beta}}{u_{j\alpha}} \tilde{3} j + O(N^{7/3+2\varepsilon}).
\]
Thus we observe that
\[
3 i j = \frac{u_{j\beta}}{u_{j\alpha}} 3 i j + O(N^{5/2+2\varepsilon}) = \frac{u_{j\beta}}{u_{j\alpha}} 3 i j + O(N^{7/3+2\varepsilon}),
\]
which proves Lemma 6.2.

7 Proof of Propositions 4.2 and 4.3

7.1 Proof of Proposition 4.2

The proof of Proposition 4.2 is divided into two steps: In the first step, we construct a set of equations for \(x\) and \(X\) from Proposition 5.1 so that solving the equation reveals a cancellation between sub-leading terms in \(5.3\). Then we prove Proposition 4.2 by simplifying each term of \(4.11\) and using the first step.

7.1.1 Expansion of remainders

In this section, we construct a vector equation for \(x\) and \(X\). Then we solve the equation to estimate the remainder term \(X\) defined below. Recall that
\[
\frac{z}{\gamma} \mathbb{E}[F'(Y)(G^2)_{aa}] = \mathbb{E}[F'(Y) a_\alpha (G^2)_{aa}] + \mathbb{E}[F'(Y) (BG^2)_{aa}] + \sqrt{7} \mathbb{E}[F'(Y)(WG^2)_{aa}] - \frac{1}{\gamma} \mathbb{E}[F'(Y)G_{aa}].
\]
Expanding the third term of \(7.1\) using Stein’s lemma with respect to \(W\), we find that
\[
\sqrt{7} \mathbb{E}[F'(Y)(WG^2)_{aa}] = -t \sqrt{m} \mathbb{E}[F'(Y)(G^2)_{aa}] - \frac{t}{N} a_\alpha - \frac{1}{\alpha} x_1 + O(N^{-1/3+\varepsilon}),
\]
where we used the fact that
\[
\left| \frac{\partial Y}{\partial X_{aa}} \right| \leq \int_{E_1} \int_{E_2} |(G^2(E + i\eta))_{aa}| dE < N^{-1/3+3\varepsilon}.
\]
Then we apply Proposition 5.1 to the second term of \(7.1\) and rearrange the equation to obtain
\[
\mathbb{E}[F'(Y)(G^2)_{aa}] = \frac{1}{N(a_\alpha - \tilde{\omega}_\alpha)^2} v_\beta^\top x - \frac{(v_\beta^\top u_\beta)^2}{u_{j\alpha}^2} \left( -\frac{1}{N(a_\alpha - \tilde{\omega}_\alpha)^2} + \frac{\tilde{m}}{N(a_\alpha - \tilde{\omega}_\alpha)^2} \right) \gamma - \frac{t}{N} \frac{1}{(a_\alpha - \tilde{\omega}_\alpha)^2} x_1 + \frac{1}{\gamma^2 (a_\alpha - \tilde{\omega}_\alpha)^2} + O(N^{-1/6+\varepsilon}),
\]
where
\[
\gamma := \left( 2 Z_{11} + \frac{\gamma}{N} \int_{E_1} \int_{E_2} \mathbb{E}[F''(Y) \mathrm{Tr} G^2 \mathrm{Im}(G^2)] dE \right).
\]
Taking the sum over \(a\) of \(7.2\) with weights \(a_{ab}^{k-1}, k = 1, 2, 3\), gives that
\[
x = u_\alpha v_\beta^\top x - \frac{(v_\beta^\top u_\beta)^2}{u_{j\alpha}^2} (-w_\alpha + \tilde{m} u_\alpha) x + t u_\alpha x_1 + \gamma^{-2} N u_\alpha + O(N^{5/6+2\varepsilon}),
\]
(7.3)
where \( w_{\alpha} \equiv (w_{\alpha 1}, w_{\alpha 2}, w_{\alpha 3})^T := \left( \int \frac{1}{(x - \omega_{\alpha})^3} d\mu_{\alpha}(x) \right) \). By symmetry, we also have

\[
x = u_{\beta} v_{\alpha}^T x - \frac{(v_{\alpha}^T u_{\beta})^2}{2u_{\beta 1}^2} (-w_{\beta} + m u_{\beta}) x + t u_{\beta} x_1 + \gamma^{-2} N u_{\beta} + O(N^{5/6 + 2\epsilon}),
\]

where \( w_{\beta} \) is defined in a similar way. Combining (7.3) and (7.4), we obtain

\[
(1 - v_{\alpha}^T u_{\beta} v_{\alpha}^T u_{\beta}) v_{\alpha}^T x = -v_{\alpha}^T u_{\beta} \frac{(v_{\alpha}^T u_{\beta})^2}{2u_{\beta 1}^2} v_{\alpha}^T (-w_{\beta} + m u_{\beta}) x - \frac{(v_{\alpha}^T u_{\beta})^2}{2u_{\beta 1}^2} v_{\alpha}^T (-w_{\alpha} + m u_{\alpha}) x + t v_{\alpha}^T u_{\alpha} v_{\beta} x_1 + t v_{\alpha}^T u_{\alpha} x_1 + \gamma^{-2} N v_{\alpha}^T u_{\alpha} v_{\beta} u_{\beta} + \gamma^{-2} N v_{\alpha}^T u_{\alpha} + O(N^{5/6 + 2\epsilon}).
\]

Applying naive bounds from Proposition 14 to (7.5) and (7.6) gives the system

\[
x = u_{\alpha} v_{\alpha}^T x + O(N^{1 + 3\epsilon}),
\]

and solving this system leads to

\[
t x = \frac{u_{\alpha} x_1}{u_{\alpha 1}} + O(N^{2/3 + 3\epsilon + x}),
\]

On the other hand, note that (7.7) implies

\[
1 - v_{\alpha}^T u_{\alpha} v_{\alpha}^T u_{\beta} = tv_{\alpha}^T u_{\alpha} u_{\beta 1} + O(N^{-2/3 + 2\epsilon}).
\]

Combining (7.5), (7.6), and (7.7), we have

\[
O(N^{5/6 + 3\epsilon}) = \left( v_{\alpha}^T u_{\alpha} \frac{(v_{\alpha}^T u_{\beta})^2}{2u_{\beta 1}^2} v_{\alpha}^T (-w_{\beta} + m u_{\beta}) + \frac{(v_{\alpha}^T u_{\beta})^2}{2u_{\beta 1}^2} v_{\alpha}^T (-w_{\alpha} + m u_{\alpha}) \right) x - \gamma^{-2} N v_{\alpha}^T u_{\alpha} v_{\beta} u_{\beta} - \gamma^{-2} N v_{\alpha}^T u_{\alpha},
\]

where we have used the fact that

\[
\frac{u_{\beta 1}}{u_{\alpha 1}} v_{\alpha}^T u_{\beta} = 1 + O(N^{-1/3 + x}) = \frac{u_{\alpha 1}}{u_{\beta 1}} v_{\beta}^T u_{\beta},
\]

which is another consequence of (7.7). Define

\[
\mathcal{K} := \frac{(v_{\alpha}^T u_{\alpha})^{5/2}}{2u_{\alpha 1}^2} v_{\beta}^T (-w_{\beta} + m u_{\beta}) + \frac{(v_{\alpha}^T u_{\beta})^{5/2}}{2u_{\beta 1}^2} v_{\alpha}^T (-w_{\alpha} + m u_{\alpha}).
\]

In what follows, we prove that \( \mathcal{K} \) is a positive real number with a constant lower bound. Note that

\[
\frac{1}{2} F''_{\alpha,0}(\tilde{\omega}_{\alpha}) \equiv \frac{1}{2} F''_{\alpha,0} = \tilde{m}^{-2} u_{\alpha 1} - \tilde{m}^{-3} u_{\alpha 1},
\]

and similar relations hold when changing the subscript \( \alpha \) to \( \beta \). Then we can represent \( \mathcal{K} \) as

\[
\mathcal{K} = -\frac{(v_{\alpha}^T u_{\alpha})^{5/2}}{2u_{\alpha 1}^2} \left( \frac{1}{2} F''_{\beta,0} + \tilde{m} v_{\beta}^T u_{\beta} (v_{\beta}^T u_{\beta} - 1) \right) - \frac{(v_{\alpha}^T u_{\beta})^{5/2}}{2u_{\beta 1}^2} \left( \frac{1}{2} F''_{\alpha,0} + \tilde{m} v_{\alpha}^T u_{\alpha} (v_{\alpha}^T u_{\alpha} - 1) \right)
\]

\[
= -\frac{(v_{\alpha}^T u_{\alpha})^{5/2}}{2u_{\alpha 1}^2} F''_{\alpha,0} - \frac{(v_{\alpha}^T u_{\beta})^{5/2}}{2u_{\beta 1}^2} F''_{\beta,0} + O(N^{-1/3 + x}),
\]

where we abbreviated \( F''_{\alpha,0} = F''_{\alpha,0}(\tilde{\omega}_{\alpha}) \) and used (7.7) and (7.8). By Lemma 3.3 and (4.1), we have \( F''_{\alpha,0}, F''_{\beta,0} < 0 \) proving that \( \mathcal{K} \) is a positive, bounded below real number. Since the second and third terms of (7.8) are real numbers, we have

\[
\text{Im} \mathcal{K} = O(N^{5/6 + \epsilon}).
\]

(7.10)
7.1.2 Proof of Proposition 4.2

We finally prove Proposition 4.2 applying above results to each term in (4.11). First of all, we calculate the time derivative of $L_+$ to simplify the first term of (4.11). By (A.20), we have

$$
\int \frac{1}{|x - \mu_0 - \tilde{m}(L_+)|^2} d\tilde{\mu}_0(x) = \frac{1}{r}.
$$

and using this we get

$$
\dot{L}_+ = \frac{\gamma L_+}{\gamma} - \gamma \tilde{m}.
$$

Next, we use the identity $\gamma HG = zG + I$ and $\dot{\gamma} = O(1)$ from Lemma 3.3 to get

$$
\dot{\gamma} E[F'(Y)(GHG)_{aa}] = \frac{\dot{\gamma}}{\gamma} E[F(x_1)G_{aa}] = \frac{\dot{\gamma}}{\gamma} \tilde{L}_+ x_1 + \frac{\dot{\gamma}}{\gamma^2} \tilde{m} - \frac{\dot{\gamma}}{\gamma^2} E[F'(Y)] + O(N^{-1/3+\delta}).
$$

By the definition of $d_1$ from (5.2), we can write

$$
E[tr G(G^2)_{aa}] = \gamma^{-1} \tilde{m} x_1 + \gamma^{-1} Z_{a11}.
$$

For the last term of (4.11), the definition of $Y$ implies that

$$
\frac{\partial Y}{\partial X_{ab}} = - \frac{1}{2} \int_{E_1} E_2 \left( G_{ba} - e_i (G^2)^i e_a \right) dE,
$$

so that

$$
\frac{1}{N^2} E[F''(Y)] \sum_{b,c} \frac{\partial Y}{\partial X_{bc}} G_{ab} G_{bc} = - \frac{1}{N^2} E[F''(Y)] \int_{E_1} E_2 (G(Im \tilde{G}^2)_{aa} dE).
$$

Plugging (7.11), (7.14) into (4.11), we finally obtain

$$
\frac{dE[F(Y)]}{dt} = \int_{E_1} E_2 \text{Im} \left[ \frac{\gamma}{2} \tilde{L}_+ x_1 - \gamma \tilde{m} x_1 - \frac{\gamma}{2} \tilde{L}_+ x_1 - N \frac{\gamma}{2} \tilde{m} E[F'(Y)] + \gamma \tilde{m} x_1 + \frac{\gamma}{2} \tilde{X} + O(N^{5/6+3\delta}) \right] dE
$$

$$
= \frac{\gamma}{2} \int_{E_1} E_2 \text{Im} \tilde{X} dE = O(N^{1/6+4\delta}),
$$

where we used (7.13) and the fact that $\tilde{m}$ is a real number. We also remark that the combined contributions of the second term of (7.13) and (7.14) is exactly $\gamma \tilde{X}/2$.

7.2 Proof of Proposition 4.3

In this section we prove Proposition 4.3. By Proposition A.1 and Theorem B.1, we find that the ESD $\mu_{H_0}$ of $H_0$ satisfies the assumptions of [20, Theorem 2.2] with high probability. Specifically, taking $\eta_0 = N^{-2/3+\chi/4}$, the diagonalization of $H_0$ is $\eta_0$-regular with high probability. Therefore [20, Theorem 2.2] implies a random version of Proposition 4.3, where random quantities $\tilde{\gamma}_{0}$ and $E_{+,t_0}$ replace $\gamma_0$ and $E_{+,t_0}$ in (4.13), respectively. These quantities are defined as follows; firstly, $E_{+,t_0}$ is the edge of $\tilde{\mu}_0 = \mu_{H_0} \triangledown \mu_{V_0}^{(t_0)}$; secondly, the scale $\tilde{\gamma}_{0}$ is defined as

$$
\tilde{\gamma}_{0} := \left( -t_0^3 \int_{E} \frac{1}{(x - E_{+,t_0} - \tilde{m}(E_{+,t_0}))^2} d\mu_{H}(x) \right)^{-1/3}.
$$

The goal of this section is to show that $|\tilde{E}_+ - E_+| = O(N^{-2/3-\chi})$ and $|\tilde{\gamma} - \gamma| = o(1)$ hold with high probability, which directly proves Proposition 4.3.

Lemma 7.1. We have

$$
|\tilde{E}_{+,t_0} - E_{+,t_0}| = O(N^{-2/3-\chi}),
$$

$$
\gamma_{0} - \tilde{\gamma}_{0} = o(1).
$$

(7.16)

(7.17)
Proof. Throughout the proof, we fix \( t = t_0 \) and omit the dependence on it. We first prove (1.10). As pointed out in [20] Equation (7.9), \( \hat{E}_+ \) is characterized as \( \hat{E}_+ = \xi_+ - t\mu_{\hat{H}}(\xi_+) \) where \( \xi_+ \) is the rightmost solution of
\[
\int_{\mathbb{R}} \frac{1}{(x - \xi)^2} d\mu_{\hat{H}} = \frac{1}{t}.
\]
By (A.21), \( E_{+t} \) and \( \xi_+ \) satisfy the same set of equations with \( \mu_{\hat{H}} \) replaced by \( \hat{\mu}_0 \). Recall also the eigenvalue rigidity for \( H_0 \), Theorem 2.6 in [5]:
\[
|\lambda_i(H_0) - \gamma_i| < N^{-2/3} \max(i, N - i)^{-1/3}, \quad \int_{\gamma_i} \tilde{m}_{\mu_0}(x) = \frac{1}{N}(i - 1/2).
\]
(7.18)
Furthermore, by (A.22) and [20] Equation (7.10), we have with high probability that
\[
\xi_+ - \sup \text{supp } \hat{\mu} \sim t^2 \sim \hat{\xi}_+ - \sup \text{supp}(\mu_{\hat{H}_0}).
\]
By Theorem 2.5 in [5], we have that
\[
|m_{\mu_0}(z) - m_{\mu_{\hat{H}_0}(z)}| < \frac{1}{N(\kappa + \eta)}.
\]
(7.19)
uniformly for all \( z = E + i\eta \in \mathcal{D} \) with \( E \geq E_+ + N^{-2/3 + \chi} \) where \( \kappa = |E - E_+| \). Let \( s = N^{-2/3 - \chi/2} \). Taking the Cauchy integral along a curve of length \( O(t^2) \) encircling \( \xi_+ + s \) gives that
\[
\left| \int_{\mathbb{R}} \frac{1}{(x - \xi_+ - s)^2} d\mu_{\hat{H}_0}(x) - \int_{\mathbb{R}} \frac{1}{(x - \xi_+ - s)^2} d\hat{\mu}_0(x) \right| \leq CN^{-1/3 + \xi t^{-4} = CN^{1/3 - 4\chi + \epsilon}}
\]
Thus we get
\[
\int_{\mathbb{R}} \frac{1}{(x - \xi_+ - s)^2} d\mu_{\hat{H}_0}(x) \leq \int_{\mathbb{R}} \frac{1}{(x - \xi_+ - s)^2} d\hat{\mu}_0(x) + C N^{1/3 - 4\chi + \epsilon}.
\]
We further have from [20] (7.23) that
\[
\int_{\mathbb{R}} \frac{1}{(x - \xi_+)^2} d\hat{\mu}_0(x) - \int_{\mathbb{R}} \frac{1}{(x - \xi_+ - s)^2} d\hat{\mu}_0(x) = s \int_{\mathbb{R}} \frac{2\xi_+ + s - 2x}{(x - \xi_+)^2(x - \xi_+ - s)^2} d\hat{\mu}_0(x) \geq \text{cst} \geq cN^{1/3 - 7\chi /2}.
\]
Therefore we have
\[
\int_{\mathbb{R}} \frac{1}{(x - \xi_+ - s)^2} d\mu_{\hat{H}_0}(x) \leq \frac{1}{t} - cN^{1/3 - 7\chi /2} + C N^{1/3 - 4\chi + \epsilon} < \frac{1}{t},
\]
by taking \( \epsilon < \chi/2 \). By a similar argument, we get
\[
\int_{\mathbb{R}} \frac{1}{(x - \xi_+ + s)^2} d\mu_{\hat{H}_0}(x) > \frac{1}{t}.
\]
Since
\[
y \rightarrow \int_{\mathbb{R}} \frac{1}{(x - \xi_+ + y)^2} d\mu_{\hat{H}_0}(x), \quad y \in (-s, s)
\]
(7.20)
is a monotone increasing function, there exists \( y_0 \in (-s, s) \) such that \( \xi_+ + y_0 = \hat{\xi}_+ \). Thus we find that
\[
E_+ - \hat{E}_+ = \xi_+ - \hat{\xi}_+ - t \int_{\mathbb{R}} \frac{1}{x - \xi_+} d\hat{\mu}_0(x) + t \int_{\mathbb{R}} \frac{1}{x - \xi_+} d\mu_{\hat{H}_0}(x)
\]
\[
= (\xi_+ - \hat{\xi}_+) \left( 1 - t \int_{\mathbb{R}} \frac{1}{x - \xi_+} d\hat{\mu}_0(x) \right) + O(N^{-2/3 - \chi})
\]
\[
= (\xi_+ - \hat{\xi}_+) \left( t \int_{\mathbb{R}} \frac{\xi_+ - \hat{\xi}_+}{x - \xi_+} d\hat{\mu}_0(x) \right) + O(N^{-2/3 - \chi})
\]
\[
= (\xi_+ - \hat{\xi}_+)^2 O(t^{-2}) + O(N^{-2/3 - \chi}) = O(N^{-2/3 - \chi}).
\]
This concludes the proof of (7.10).
Next, we turn to the proof of (7.17). Since \( \gamma_t, \tilde{\gamma}_t \sim 1 \) from (3.6) and [20 Lemma 2.3], we get

\[
|\gamma_t - \tilde{\gamma}_t| \leq Ct^3 \left| \int \frac{1}{(x - \xi_t)^3} \mu_0(x) - \int \frac{1}{(x - \xi_t)^3} \mu_{\alpha}(x) \right|
\]

\[
\leq Ct^3 \left| \int \frac{1}{(x - \xi_t)^3} \mu_0(x) - \frac{1}{(x - \xi_t)^3} \mu_{\alpha}(x) \right| + Ct^3 \left| \int \frac{1}{(x - \xi_t)^3} \mu_{\alpha}(x) \right|.
\]

(7.21)

The second term on the RHS of (7.21) can be estimated using (7.19), so that taking the Cauchy integral gives that

\[
t^3 \left| \int \frac{1}{(x - \xi_t)^3} \mu_0(x) - \int \frac{1}{(x - \xi_t)^3} \mu_{\alpha}(x) \right| \leq N^{1+\epsilon} = N^{-3\kappa + \epsilon}.
\]

The first term on the RHS of (7.21) can be estimated as

\[
t^3 \left| \int \frac{1}{(x - \xi_t)^3} \mu_0(x) - \int \frac{1}{(x - \xi_t)^3} \mu_{\alpha}(x) \right| \leq O(t^{-2}(|\xi_t - \hat{\xi}_t|) \leq N^{-5\kappa/2},
\]

where we used \(|\xi_t - \hat{\xi}_t| \leq N^{-2/3-\kappa/2} in the last inequality. This completes the proof of Lemma 7.1.

By Lemma 7.1 and the arguments at the beginning of Section 7.2, we have proved Proposition 4.3.

Acknowledgements

The authors would like to thank Ji Oon Lee for helpful discussions. The work of J. Park was partially supported by National Research Foundation of Korea under grant number NRF-2019R1A5A1028324. The work of H. C. Ji was partially supported by ERC Advanced Grant "RMTBeyond" No. 101020331.

A Stability of \( \mu_A \boxplus \mu_B \boxplus \mu_{sc}^{(t)} \) and \( \mu_\alpha \boxplus \mu_\beta \boxplus \mu_{sc}^{(t)} \)

Recall that we omitted the subscript \( t \) to denote, say, \( \omega_{A,t} \) by \( \omega_A \). The two goals of this section are to prove that the system of equations \( \Phi_{\alpha,\beta} = 0 \) defined in (3.8) is stable around \( z = E_+ \) and to extend the same result to \( \Phi_{AB} = 0 \) by comparison. First, we introduce notations for quantities that are widely used throughout the paper.

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

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

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

Similarly we define \( S_{AB}, T_{\alpha}, \) and \( T_B \) to be the same quantities with \( (\alpha, \beta) \) replaced by \( (A, B) \). From (3.8), we see that the edge \( E_+ \) satisfies \( S_{\alpha,\beta}(E_+) = 0 \). The main result of this section is the following proposition, whose proof is postponed to the end of this section:

**Proposition A.1.** Let \( \sigma > 0 \) be fixed. Then there exist constants \( \tau > 0 \) and \( (large) N_0 \in \mathbb{N} \) such that each of the following holds uniformly over \( z \in D_r(N^{-1+\sigma}, 1) \) and \( t \in [0, 1] \) for all \( N \geq N_0 \).

(i) There exist positive constants \( k \) and \( K \) such that

\[
\min |a_i - \omega_A(z)| \geq k, \quad \min |b_i - \omega_B(z)| \geq k, \quad |\omega_A(z)| \leq K, \quad |\omega_B(z)| \leq K.
\]

(ii) Recall that \( \hat{\mu}_t \) denotes the free convolution \( \mu_A \boxplus \mu_B \boxplus \mu_{sc}^{(t)} \). For its Stieltjes transform \( m_{\hat{\mu}_t} \), we have

\[
\text{Im } m_{\hat{\mu}_t}(z) \sim \begin{cases} \sqrt{\kappa + \eta}, & \text{if } E \in \text{supp } \hat{\mu}_t, \\ \sqrt{\kappa - \eta}, & \text{if } E \notin \text{supp } \hat{\mu}_t, \end{cases}
\]

where we denoted \( z = E + i\eta \) and \( \kappa = |E_+ - E| \).
(iii) There exists a constant $C > 0$ such that

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

Moreover, there exist positive constants $\delta$ and $c$ such that, whenever $|z - E_+| \leq \delta$,

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

(iv) There exists a constant $C > 0$ such that

$$|\omega_A(z)| \leq C \frac{1}{\sqrt{\kappa + \eta}}, \quad |\omega_B(z)| \leq C \frac{1}{\sqrt{\kappa + \eta}}, \quad |S'_{AB}(z)| \leq C \frac{1}{\sqrt{\kappa + \eta}}.$$

A.1 Stability of $\mu_\alpha \boxplus \mu_\beta \boxplus \mu^{(1)}_{sc}$

In this subsection, we study regularity properties of $\mu_\alpha \boxplus \mu_\beta \boxplus \mu^{(1)}_{sc}$. Also, we present the proof of Lemma 3.3 at the end of this subsection.

**Lemma A.2.** Let $\mu_\alpha$ and $\mu_\beta$ be probability measures in Definition 2.2. Then for each $t \in [0, 1]$ there exist unique Borel measures $\tilde{\mu}_{\alpha,t}$ and $\tilde{\mu}_{\beta,t}$ on $\mathbb{R}$ such that

$$F_{\alpha,t}(z) - z = m_{\alpha,t}(z), \quad F_{\beta,t}(z) - z = m_{\beta,t}(z) \quad (A.1)$$

for all $z \in \mathbb{C}_+$. Furthermore, we have that

$$\tilde{\mu}_{\alpha,0}(\mathbb{R}) = \int_{\mathbb{R}} x^2 d\mu_\alpha(x) - \left( \int_{\mathbb{R}} x d\mu_\alpha(x) \right)^2, \quad [E_+^\alpha - \tau, E_+^\alpha] \subset \text{supp} \tilde{\mu}_{\alpha,0} \subset [E_+^\alpha, E_+^\alpha], \quad (A.2)$$

$$\tilde{\mu}_{\beta,0}(\mathbb{R}) = \int_{\mathbb{R}} x^2 d\mu_\beta(x) - \left( \int_{\mathbb{R}} x d\mu_\beta(x) \right)^2, \quad [E_+^\beta - \tau, E_+^\beta] \subset \text{supp} \tilde{\mu}_{\beta,0} \subset [E_+^\beta, E_+^\beta],$$

and for all $t \in [0, 1]$ that

$$\tilde{\mu}_{\alpha,t} = \tilde{\mu}_{\alpha,0} + t\mu_\alpha, \quad \tilde{\mu}_{\beta,t} = \tilde{\mu}_{\beta,0} + t\mu_\beta. \quad (A.3)$$

**Proof.** The proof is a minor modification of that of [8, Lemma 3.5], and we sketch its proof here for readers’ convenience. We prove the result only for $\mu_\alpha$ and that for $\mu_\beta$ is exactly the same.

The existence and uniqueness of $\tilde{\mu}_\alpha$ follow from Nevanlinna-Pick representation theorem, and the formula for $\tilde{\mu}_\alpha(\mathbb{R})$ is a direct consequence of (A.1) and the definition of $F_\alpha$ in (3.3). Given the uniqueness, we see from (A.3) that $\tilde{\mu}_{\alpha,t} = \tilde{\mu}_{\alpha,0} + t\mu_\alpha$.

In order to prove $\text{supp} \tilde{\mu}_{\alpha,0} \subset [E_+^\alpha, E_+^\alpha]$ and observe that for each $x \in [E_+^\alpha, E_+^\alpha]$ that

$$\lim_{y \to 0^+} \text{Im } m_{\mu_{\alpha,0}}(x + iy) = \lim_{y \to 0^+} \text{Im } F_{\alpha,0}(x + iy) = |m_{\mu_{\alpha}}(x)|^{-2} \lim_{y \to 0^+} \text{Im } m_{\mu_{\alpha}}(x + iy) = 0.$$  

Then Stieltjes inversion directly implies $[E_+^\alpha, E_+^\alpha] \subset \text{supp} \tilde{\mu}_{\alpha,0}$ as desired.

Finally we prove the inclusion $[E_+^\alpha - \tau, E_+^\alpha] \subset \text{supp} \tilde{\mu}_{\alpha,0}$. Suppose on the contrary that there exists a nonempty open interval $I \subset (E_+^\alpha - \tau, E_+^\alpha) \setminus \text{supp} \tilde{\mu}_{\alpha,0}$. Since $I \subset \text{supp} \tilde{\mu}_{\alpha,0}$, the function $z \mapsto F_{\alpha,0}(z) - z$ extends analytically through $I$ via Schwarz reflection which satisfies $F_{\alpha,0}(x) - x \in \mathbb{R}$ for each $x \in I$. Then this leads to a meromorphic extension of $m_{\mu_{\alpha}}$ by the definition of $F_{\alpha,0}$. This extension must satisfy $\text{Im } m_{\mu_{\alpha}}(x) = 0$ for almost all $x \in I$, which contradicts Definition 2.2.

We introduce the following result from [12] which gives a necessary condition that a free additive convolution has unbounded Stieltjes transforms:

**Lemma A.3** (Theorem 7 of [12]). Let $\mu$ and $\nu$ be compactly supported Borel probability measures on $\mathbb{R}$. If the image $m_{\mu \boxplus \nu}(\mathbb{C}_+)$ is unbounded, then there exist real numbers $u$ and $v$ such that $\mu(\{u\}) + \nu(\{v\}) > 1$.

Clearly our measures $\mu_\alpha$ and $\mu_\beta$ from Definition 2.2 are compactly supported and since they are absolutely continuous it also follows that $\mu_\alpha(\{u\}) + \mu_\beta(\{v\}) = 0$ for all choices of $u$ and $v$. Therefore it follows that $m_{\mu_\alpha}$ is bounded on $\mathbb{C}_+$ by Lemma A.3 and the same bound applies to $m_{\mu_\beta}(z) = m_{\mu_\alpha}(z + t\mu_\alpha)$. One advantage of applying this result is that we can bypass the assumption in [8] that $\sup_{z \in \mathbb{C}_+} |m_{\mu_{\alpha}}(z)| \leq C$.  

28
Lemma A.4. The maps \( \omega_{\alpha,t}(z), \omega_{\beta,t}(z), \) and \( m_{\mu_t}(z) \) are continuous in \( (t, z) \in [0, \infty) \times (\mathbb{C}_+ \cup \mathbb{R}) \).

Proof. Recall that \( \omega_{\alpha,0}(z) = \omega_{\alpha,0}(z + \operatorname{Im} z) \) and \( \omega_{\beta,0}(z) = \omega_{\beta,0}(z + \operatorname{Im} z) \). Since \( \omega_{\alpha,0} \) and \( \omega_{\beta,0} \) continuously extend to \( \mathbb{C}_+ \cup \mathbb{R} \) by [11, Theorem 3.3], it suffices to consider only \( m_{\mu_t}(z) \). Thus, our goal here is to prove the following statement; for all fixed \( \epsilon > 0, z \in \mathbb{C}_+, \) and \( t \in [0, \infty) \), there exists \( \delta > 0 \) such that \( |z - w| < \delta \) and \( |t - s| < \delta \) imply
\[
|m_{\mu_t}(z) - m_{\mu_s}(w)| < \epsilon. \tag{A.4}
\]

First, we prove that \( m_t(z) \) is continuous at \( (0, z) \) for each fixed \( z \in \mathbb{C}_+ \). Take \( \delta > 0 \) so that \( |w - z| < 3\delta \) implies \( |m_{\mu_0}(z) - m_{\mu_0}(w)| < \epsilon/100 \). Now we take \( \delta_1 \) to satisfy \( \delta_1 \sup_{z \in \mathbb{C}_+} |m_{\mu_0}(z)| < \delta \). Then, for all \( t \in (0, \delta_1) \) and \( w \in \mathbb{C}_+ \) with \( |z - w| < \delta \), we have
\[
|m_{\mu_t}(z) - m_{\mu_0}(w)| = |m_{\mu_t}(z) - m_{\mu_0}(w + \operatorname{Im} z)| < \epsilon.
\]

Next, we prove the continuity at \( (t, z) \in (0, \infty) \times \mathbb{C}_+ \). It is clear that the result follows from the following assertion; there exists a constant \( C > 0 \) depending on \( R > 0 \) such that the following holds whenever \( w \in \mathbb{C}_+ \) and \( t, s > 0 \);
\[
|tm_{\mu_t}(w) - sm_{\mu_s}(w)| \leq C \min(t, s)^{-1/3}((\max(t, s) + |w|)^4 |t - s|^{1/3}. \tag{A.5}
\]

In what follows, we focus on proving (A.5). We suppose \( s > t \) without loss of generality and write
\[
m_t = m_{\mu_t}(w), \quad m_s = m_{\mu_s}(w), \quad \xi_t = w + tm_{\mu_t}(w), \quad \xi_s = w + sm_{\mu_s}(w)
\]
to simplify the presentation. Using the equation \( m_t = m_{\mu_0}(\xi_t) \), we find that
\[
m_t - m_s = m_{\mu_0}(\xi_t) - m_{\mu_0}(\xi_s) = (\xi_t - \xi_s) \int_{\mathbb{R}} \frac{1}{(x - \xi_t)(x - \xi_s)} \, d\mu_0(x).
\]
Using the definition of \( \xi_t \) and rearranging the equation, we get
\[
(\xi_t - \xi_s) \left( 1 - t \int_{\mathbb{R}} \frac{1}{(x - \xi_t)(x - \xi_s)} \, d\mu_0(x) \right) = (t - s)m_s,
\]
and we will derive (A.5) from a lower bound for the left-hand side of (A.6). Using the fact that
\[
\operatorname{Im} \xi_s = \operatorname{Im} w + t \operatorname{Im} m = \operatorname{Im} w + s \operatorname{Im} \xi_s \int_{\mathbb{R}} \frac{1}{|x - \xi_s|^2} \, d\mu_0(x),
\]
and \( \xi_t, \xi_s \in \mathbb{C}_+ \) and \( |\xi_t|, |\xi_s| \leq |w| + sC \) in the last line. Now plugging (A.8) into (A.6), we get
\[
C(t + |w|)^{-4}|\xi_t - \xi_s|^3 \leq C|t - s|,
\]
which implies (A.5).
\[\square\]

Lemma A.5. There exists a positive constant \( \tau \) such that for all fixed \( s, \eta_M > 0 \) there is \( C > 0 \) with
\[
\sup_{t \in [0, s]} \sup_{z \in \mathbb{D}_+(0, \eta_M)} |\omega_{\alpha,t}(z)| + |\omega_{\beta,t}(z)| \leq C.
\]

29
Proof. We prove the bound for $\omega_{\alpha,t}$ and the same proof applies to $\omega_{\beta,t}$. First of all, from (2.2) we find that for each fixed $M > 0$ there exists a constant $c > 0$ such that
\[
\inf \{ |m_{\beta}(\omega)| : \Re \omega \in (E_{\beta}^+, \tau_\beta, E_{\beta}^+ + M), \Im \omega \in (0, 2\eta_M) \} \geq c.
\] (A.9)
On the other hand by Lemma [A.2] there exist constants $C_1, C_2 > 0$ such that
\[
|F_{\alpha,t}(\omega)| \geq \frac{1}{2}\omega\quad \text{and} \quad |F_{\alpha,t}(\omega) - \omega| \leq C_2|\omega|^{-1}
\]
whenever $|\omega| \geq C_1$.
Now we assume in contrary that $|\omega_{\alpha,t}(z)| \geq K$ for some $z \in \mathcal{D}(0, \eta_M)$ with $\tau < \tau_\beta$ and $K > 0$ to be chosen later. In particular if $K > C_1$ we have
\[
|\omega_{\beta,t}(z) - z| = |F_{\alpha,t}(\omega_{\alpha,t}(z)) - \omega_{\alpha,t}(z)| \leq C_2K^{-1}.
\]
Thus we can take $K$ to be large enough so that
\[
\Re \omega_{\beta,t}(z) \geq E_{\beta,t} - \tau - C_2K^{-1} > E_{\beta}^+ - \tau_\beta, \quad \Im \omega_{\beta,t}(z) < \eta_M + C_2K^{-1} < 2\eta_M,
\]
where we used the fact that $E_{\beta}^+ \leq E_{\beta,t} \leq E_{\beta}^+ + E_{\alpha}^+ + 2\sqrt{7}$ from [33] Lemma 3.1. In other words, $\omega_{\beta,t}(z)$ lies within the domain in (A.9). Then we obtain
\[
|F_{\beta,t}(\omega_{\beta,t}(z))| \leq \frac{1}{|m_{\mu,\beta}(\omega_{\beta,t}(z))|} + s|m_{\mu,\beta}(z + tm_{\mu,\beta}(z))| \leq C
\]
for some constant $C > 0$. After raising $K$ further, we have a contradiction since
\[
|F_{\beta,t}(\omega_{\beta,t}(z))| = |F_{\alpha,t}(\omega_{\alpha,t}(z))| \geq \frac{|\omega_{\alpha,t}(z)|}{2}.
\]
This proves $|\omega_{\alpha,t}(z)| \leq K$, and the bound for $\omega_{\beta,t}(z)$ follows from the same proof.

Lemma A.6. Recall that $E_{\beta,t} = \sup \text{supp } \mu_t$. For each $t \geq 0$, the maps $\omega_{\alpha,t}$ and $\omega_{\beta,t}$ are real-valued and monotone increasing on $(E_{\alpha,t}, \infty)$, and they map into $(E_{\alpha}^+, \infty)$ and $(E_{\beta}^+, \infty)$, respectively.

Proof. For any $z \in [E_{\alpha,t}, \infty)$ we have from (3.3) that
\[
\Im \omega_{\alpha,t}(z) + \Im \omega_{\beta,t}(z) = \Im F_{\mu}(z) + \Im z = 0,
\]
which implies $\Im \omega_{\alpha,t}(z) = \Im \omega_{\beta,t}(z) = 0$ since both of them should be nonnegative. Furthermore, for a large enough positive $z_0$, we have $\omega_{\alpha,t}(z_0) = m_{\mu,\alpha}^{-1}(m_{\mu}(z_0)) \in (E_{\alpha}^+, \infty)$. Now we suppose on the contrary that there exists $z \in [E_{\beta,t}, \infty)$ such that $\omega_{\beta,t}(z) < E_{\alpha}^+$. Then, since $\Re \omega_{\alpha,t}(z)$ is a continuous real function, there must be another point $w$ between $z$ and $z_0$ such that $\Re \omega_{\alpha,t}(w) \in (E_{\alpha}^+, \tau_\alpha, E_{\alpha}^+)$. Then by (2.2) we have that
\[
\lim_{y \to 0} \Im m_{\mu}(w + iy) = \lim_{y \to 0} \Im m_{\mu,\alpha}(\omega_{\alpha,t}(w + iy)) = \lim_{y \to 0} \Im m_{\mu,\alpha}(\omega_{\alpha,t}(w) + iy) > 0,
\]
which contradicts $w \notin \text{supp } \mu_t$. Here we used Lindelöf’s theorem [11] Theorem 2.7] in the second equality when $\Im \omega_{\alpha,t}(w) = 0$. Thus we have proved that $\omega_{\alpha,t}$ maps $[E_{\alpha,t}, \infty)$ into $[E_{\alpha}^+, \infty)$. The fact that $\omega_{\alpha,t}$ is increasing follows directly from chain rule as in [8] Lemma 3.3.

Lemma A.7. There exist positive constants $\tau$ and $k_0$ such that
\[
\inf_{t \in [0, 1]} \inf_{z \in \mathcal{D}(0, \infty)} \inf_{x \in \text{supp } \mu_{\alpha}} |\omega_{\alpha,t}(z) - x| \geq k_0, \quad \inf_{t \in [0, 1]} \inf_{z \in \mathcal{D}(0, \infty)} \inf_{x \in \text{supp } \mu_{\beta}} |\omega_{\beta,t}(z) - x| \geq k_0.
\]

Proof. We only present an outline of the proof since it is a minor modification of [8] Lemma 3.7. First of all, we prove that the following statement implies the result; there exists a constant $k > 0$ such that
\[
\inf_{t \in [0, 1]} |\omega_{\alpha,t}(E_{\alpha}^+) - E_{\alpha}^+| > k, \quad \inf_{t \in [0, 1]} |\omega_{\beta,t}(E_{\beta}^+) - E_{\beta}^+| > k.
\] (A.10)
Assuming \((A.10)\), we find from Lemma \([A.4]\) that for a sufficiently small \(\tau\)
\[
\inf_{t \in [0, 1]} \inf_{x \in D_{\tau}(0, \tau)} \inf_{x \in \text{supp } \mu_{\alpha}} \inf_{x \in \text{supp } \mu_{\beta}} |\omega_{\alpha,t}(E_{+}, t) - x| > k, \quad \inf_{t \in [0, 1]} \inf_{x \in D_{\tau}(0, \tau)} \inf_{x \in \text{supp } \mu_{\beta}} |\omega_{\beta,t}(E_{+}, t) - x| > k.
\]
Since \(\text{Im } \omega_{\beta,t}(z) = \text{Im } \omega_{\beta,0}(z + t \mu_{\beta}(z)) \geq \text{Im } z\), the result directly extends to \(D_{\tau}(0, \infty)\).
In order to prove \((A.10)\), we recall the following identities from \((3.3)\);
\[
\text{Im } m_{\mu_{\alpha}}(z) = \text{Im } \omega_{\alpha,t}(z) \int_{\mathbb{R}} \frac{1}{|x - \omega_{\alpha,t}(z)|^2} d\mu_{\alpha}(x) = \text{Im } \omega_{\beta,t}(z) \int_{\mathbb{R}} \frac{1}{|x - \omega_{\beta,t}(z)|^2} d\mu_{\beta}(x),
\]
\[
\text{Im } \omega_{\alpha,t}(z) + \text{Im } \omega_{\beta,t}(z) - \text{Im } z = \text{Im } F_{\mu}(z) = \frac{\text{Im } m_{\mu_{\alpha}}(z)}{|m_{\mu_{\alpha}}(z)|^2} + t \text{Im } m_{\mu_{\beta}}(z).
\]
We then have
\[
R_{\alpha}(\omega_{\alpha,t}(z)) + R_{\beta}(\omega_{\beta,t}(z)) - 1 = \frac{|m_{\mu_{\alpha}}(z)|^2}{|m_{\mu_{\alpha}}(z)|^2} + t|m_{\mu_{\beta}}(z)|^2 \geq 0, \quad (A.11)
\]
where we defined for \(\omega \in \mathbb{C}_{+}\)
\[
R_{\alpha}(\omega) := \left| \int_{\mathbb{R}} \frac{1}{x - \omega} d\mu_{\alpha}(x) \right|^2 / \int_{\mathbb{R}} \frac{1}{|x - \omega|^2} d\mu_{\alpha}(x)
\]
and \(R_{\beta}(\omega)\) analogously.
We now prove \((A.10)\) for \(\omega_{\alpha,t}\), and the result for \(\omega_{\beta,t}\) follows by symmetry. Proceeding as in the proof of \([S] \text{ Lemma 3.7}\), we find that \(\omega_{\alpha,t}(E_{+}, t) \geq E_{\alpha}^+\) and \(\omega_{\beta,t}(E_{+}, t) \geq E_{\beta}^+\) imply
\[
R_{\alpha}(\omega_{\alpha,t}(E_{+}, t)) \leq C_{\alpha}(1^{1/|t_{\alpha}|_{\infty}}, \quad R_{\beta}(\omega_{\beta,t}(E_{+}, t)) \leq C_{\beta}(1^{1/|t_{\beta}|_{\infty}}, \quad (A.12)
\]
for a constant \(C_{\alpha}\) independent of \(t\). On the other hand, due to Cauchy-Schwarz inequality we have
\[
R_{\beta}(\omega) \leq 1 - c_{\beta}(\omega) \quad \omega \in (E_{\beta}^+, \infty), \quad (A.13)
\]
for a positive continuous function \(c_{\beta}\) on \((E_{\beta}^+, \infty)\). Combining Lemma \([A.5] \text{ A.12}\), and \((A.13)\) implies that
\[
\sup_{t \in [0, 1]} R_{\beta}(\omega_{\beta,t}(E_{+}, t)) \leq 1 - c
\]
for a constant \(c > 0\). Therefore by \((A.11)\) we have
\[
\omega_{\alpha,t}(E_{+}, t) - E_{\alpha}^+ \geq C_{\alpha} t_{\alpha}^{1/(1 - |t_{\alpha}|_{\infty})}.
\]
This concludes the proof of \((A.10)\) and thus Lemma \([A.7]\).

Lemma A.8. For all \(z \in \mathbb{C}_{+}\) and \(t \in [0, 1]\) we have
\[
|\left( F_{\alpha,t}'(\omega_{\alpha,t}(z)) - 1 \right) \left( F_{\beta,t}'(\omega_{\beta,t}(z)) - 1 \right) | \leq 1, \quad (A.14)
\]
and the upper edge \(z = E_{+}, t\) is the largest real point at which equality hold. Furthermore, we have
\[
\left( F_{\alpha,t}'(\omega_{\alpha,t}(E_{+}, t)) - 1 \right) \left( F_{\beta,t}'(\omega_{\beta,t}(E_{+}, t)) - 1 \right) = 1. \quad (A.15)
\]
Proof. The proof that \(E_{+}\) satisfies \((A.15)\) is identical to that of Lemma 3.8 in \([S]\). To prove the remaining part, that \(E_{+}\) is the largest such point, we observe from Lemma \([A.2]\) that the LHS of \((A.14)\) decreases as \(\omega_{\alpha,t} > E_{\alpha}^+\) and \(\omega_{\beta,t} > E_{\beta}^+\) increase. Since \(\omega_{\alpha}||E_{+}, \infty)\) and \(\omega_{\beta}||E_{+}, \infty)\) are increasing real functions mapping into \((E_{\alpha}^+ + k, \infty)\) and \((E_{\beta}^+ + k, \infty)\), the result follows.

Proposition A.9. For each \(t \in [0, 1]\) there exist positive \(\gamma_{\alpha,t}\) and \(\gamma_{\beta,t}\) such that the following hold uniformly over \(t \in [0, 1]\) and \(z \in D_{\tau}(0, \eta_{\alpha})\);
\[
\omega_{\alpha,t}(z) - \omega_{\alpha,t}(E_{+}, t) = \gamma_{\alpha} \sqrt{z - E_{+}, t} + O(|z - E_{+}, t|),
\]
\[
\omega_{\beta,t}(z) - \omega_{\beta,t}(E_{+}, t) = \gamma_{\beta} \sqrt{z - E_{+}, t} + O(|z - E_{+}, t|),
\]
\[
\gamma_{\alpha,t} \sim 1 \sim \gamma_{\beta,t}.
\]
Proof. Given Lemmas [A.4] [A.5] [A.8], the proof is almost identical to that of [S] Lemma 3.8 except some minor changes to make the result uniform over $t$. We present below how we modify their proof.

Note the map $F_{\alpha,t}$ has an inverse in a neighborhood of $F_{\alpha,t}(\omega_{\alpha,t}(E_{+},t)) = F_{\beta,t}(\omega_{\beta,t}(E_{+},t))$ that maps into a neighborhood of $\omega_{\alpha,t}(E_{+},t)$. The first modification is to show that both the domain and image of the inverse $F_{\alpha,t}^{-1}$ can have size of $O(1)$. Note that

$$F_{\alpha,t}(\omega_1) - F_{\alpha,t}(\omega_2) = (\omega_1 - \omega_2) \left( 1 + \frac{1}{\int_{\mathbb{R}} \frac{1}{|x - \omega_1| |x - \omega_2|} d\mu_{\alpha,t}(x) } \right).$$

When $\Re \omega_i \in (\mathcal{E}_n^+ + k, \infty), \Im \omega_i \in (0, k)$ for $i = 1, 2$ and a positive constant $k$, we have

$$\Re \frac{1}{(x - \omega_1)(x - \omega_2)} = \frac{(x - Re \omega_1)(x - Re \omega_2) - \Im \omega_1 \Im \omega_2}{|x - \omega_1|^2 |x - \omega_2|^2} > 0, \quad x \in \operatorname{supp} \mu_{\alpha,t}$$

for all $x \in \operatorname{supp} \mu_{\alpha,t}$, so that

$$\left| \frac{F_{\alpha,t}(\omega_1) - F_{\alpha,t}(\omega_2)}{\omega_1 - \omega_2} \right| > 1. \quad (A.16)$$

Thus $F_{\alpha,t}$ restricted to the domain $D_\alpha := \{z : \Re z > E_\alpha^+ + k_0/2, \Im z \in (0, k_0/2)\}$ has an analytic inverse by the open mapping theorem, where $k_0$ is from Lemma [A.7].

We next prove that the image $F_{\alpha,t}(D_\alpha)$ contains a disk around $F_{\alpha,t}(\omega_{\alpha,t}(E_{+},t))$ whose radius admits a uniform lower bound over $t \in [0, 1]$. First we observe that we have an upper bound for $|\omega - \omega_{\alpha,t}(E_{+},t)| = k_0/3$ implies

$$|F_{\alpha,t}(\omega) - F_{\alpha,t}(\omega_{\alpha,t}(E_{+},t))| > k_0/3.$$ 

Now taking $x$ so that $|x - F_{\alpha,t}(\omega_{\alpha,t}(E_{+},t))| < k_0/6$, we have

$$\frac{1}{|x - F_{\alpha,t}(\omega)|} < 6/k_0, \quad |\omega - \omega_{\alpha,t}(E_{+},t)| = k_0/3.$$ 

If $x - F_{\alpha,t}(\omega)$ has no zero in the domain $|\omega - \omega_{\alpha,t}(E_{+},t)| < k_0/3$, we have a contradiction from the maximum modulus principle. This proves $x = F_{\alpha,t}(\omega)$, so that $x \in F_{\alpha,t}(D_\alpha)$.

As in [S] Lemma 3.8, we define

$$\tilde{z}(\omega) := -F_{\beta,t}(\omega) + \omega + F_{\alpha,t}^{-1} \circ F_{\beta,t}(\omega).$$

By the argument above, we find that $\tilde{z}$ is indeed an analytic function in domain $|\omega - \omega_{\beta,t}(E_{+},t)| < k_0/6C_\beta$. Now we follow the lines of [S] Lemma 3.8 to get $\tilde{z}'(\omega_{\beta,t}(E_{+},t)) = 0$ from Lemma [A.3] so that

$$\left| \tilde{z}(\omega) - E_{+},t - \frac{1}{2} \tilde{z}''(\omega_{\beta,t}(E_{+},t))(\omega - \omega_{\beta,t}(E_{+},t))^2 \right| \leq C|\omega - \omega_{\beta,t}(E_{+},t)|^3, \quad (A.17)$$

where the constant $C$ is chosen uniformly over $t \in [0, 1]$ and $|\omega - \omega_{\beta,t}(E_{+},t)| < k_0/6$.

We apply Lemma [A.4] to $\omega_{\beta,t}$, so that there exists $\tau > 0$ such that $|z - E_{+},t| < \tau$ implies $|\omega_{\beta,t}(z) - \omega_{\beta,t}(E_{+},t)| < k_0/6$. Then (A.17) reads

$$|z - E_{+},t - \gamma_{\beta,t}(\omega - \omega_{\beta,t}(E_{+},t))^2| \leq \frac{1}{2} C|\omega_{\beta,t}(z) - \omega_{\beta,t}(E_{+},t)|^3, \quad |z - E_{+},t| < \tau, \quad (A.18)$$

where we defined

$$\gamma_{\beta,t} = \sqrt{\frac{2}{\tilde{z}''(\omega_{\beta,t}(E_{+},t))}}.$$ 

We again follow [S] Lemma 3.8 to get $\gamma_{\beta,t} \sim 1$ uniformly over $t$. Inverting the expansion (A.18) (taking smaller $\tau$ if necessary) concludes the proof of Proposition [A.9].

Now that we have established Proposition [A.9], the following result can be easily proved:
Corollary A.10. The following hold uniformly over \( t \in [0, 1] \) and \( z \in \mathbb{C}_+ \cap \{ z : |z - E_{+, t}| \leq \tau \} \):

\[
\begin{align*}
  m'_{\mu}(z) &\sim |z - E_{+, t}|^{-1/2}, & m''_{\mu}(z) &\sim |z - E_{+, t}|^{-3/2}, \\
  \omega'_{\alpha}(z) &\sim |z - E_{+, t}|^{-1/2}, & \omega''_{\alpha}(z) &\sim |z - E_{+, t}|^{-3/2}, \\
  F'_{\alpha,t}(\omega_{\alpha,t}(z)) &\sim 1, & F''_{\alpha,t}(\omega_{\alpha,t}(z)) &\sim 1, & \mathcal{T}_\alpha(z) &\sim 1.
\end{align*}
\]

Furthermore, we have the following in the larger domain \( D_T(0, \eta_M) \):

\[
\begin{align*}
  \text{Im} m'_{\mu}(z) \sim \text{Im} \omega_{\alpha}(z) \sim \text{Im} \omega_{\beta}(z) &\sim \begin{cases} \sqrt{\kappa + \eta}, & \text{if } E \leq E_{+, t}; \\
\sqrt{\kappa - \eta}, & \text{if } E > E_{+, t}; \end{cases} \int \sqrt{|x - \omega_{\alpha}(E_{+, t})|^2} |\mu_0(x)| \sim 1.
\end{align*}
\]

\[\mathcal{T}_\alpha(z) \sim \mathcal{T}_\beta(z).\]  

Proof. The proof of Corollary A.10 is the same as [8, Corollaries 3.10 and 3.11] except some minor modifications. For example, when proving (A.19) for the regime \( \kappa \sim 1 \), we used the following fact:

\[
\inf\{\omega_{\alpha,t}(E) - \omega_{\alpha,t}(E_{+, t}) : E > E_{+, t} + \kappa, t \in [0, 1]\} > c(\kappa)
\]

for some constant \( c(\kappa) > 0 \) depending only on \( \kappa > 0 \). This inequality is a direct consequence of Lemmas A.4 and A.6. Similarly we can easily modify the proof of [8] to prove Corollary A.10, and we omit further details.

Proof of Lemma A.6. The first, second, third, and fourth parts of the lemma are proved in Lemmas A.4, A.5, A.7, and A.8. Also, (3.6) is a direct consequence of (3.2) and Proposition A.9. In particular, we have

\[\rho(E) = \text{Im} \omega_{\alpha}(E) \int \frac{1}{|x - \omega_{\alpha}(E_{+, t})|^2} |\mu_0(x)|, \quad E \in [E_{+, -}, E_{+, +}],\]

so that

\[\gamma_t^{3/2} = \gamma_{\alpha,t} \int \frac{1}{|x - \omega_{\alpha}(E_{+, t})|^2} |\mu_0(x)| \sim 1.\]

Thus it only remains to prove the last assertion that \( \frac{d}{dt} \gamma_t \sim \sqrt{t} \). Since both of \( \mu_0 \) and \( \mu_{\alpha}(t) \) are Jacobi-type measures for each fixed \( t > 0 \), a direct application of [7, Proposition 4.7] gives that the upper edge \( E_{+, t} \) is the rightmost solution \( z = E_{+, t} \) of the equation

\[\int \frac{1}{|x - z - tm_{\mu}(z)|^2} |\mu_0(x)| = \frac{1}{t}.\]

That is, denoting \( \xi_t = E_{+, t} + tm_{\mu}(E_{+, t}) \),

\[\int \frac{1}{|x - \xi_t|^2} |\mu_0(x)| = \frac{1}{t},\]

and we have \( E_{+, t} = \xi_t - tm_{\mu_0}(\xi_t) \). Using the fact that \( \mu_0 \) has square-root decay, we have

\[\int \frac{1}{|x - \xi_t|^2} |\mu_0(x)| \sim (|\xi_t - E_{+, 0}|^{-1/2},
\]

so that

\[ct^2 \leq \xi_t - E_{+, 0} \leq Ct^2, \quad \gamma_t \leq E_{+, t} - E_{+, 0} \leq Ct\]

for some fixed constants \( c, C > 0 \), where we used \( E_{+, t} = \xi_t - tm_{\mu_0}(\xi_t) \). On the other hand, by the same proposition in [7] (see also [20, Lemma 2.3]) we have that

\[\gamma_t = \left( -t^3 \int_R \frac{1}{|x - \xi_t|^3} \right)^{-1/3} \approx \gamma_t \sim 1.\]

Note that since \( \gamma_t \sim 1 \) it suffices to prove that \( \frac{d}{dt} \gamma_t \sim 1 \). This follows from a straight forward calculation using (A.22), together with the fact that

\[\frac{d}{dt} \xi_t = -\frac{1}{t^2} \left( \int_R \frac{1}{|x - \xi_t|^3} |\mu_0(x)| \right)^{-1} \sim t.\]
A.2 Stability of $\mu_A \boxplus \mu_B \boxplus \mu_{sc}$

In this section, we prove Proposition [A.14] following the proof of [S] Proposition 3.1. More specifically, we prove an upper bound for the distance between $(\omega_{A,t}(z), \omega_{B,t}(z))$ and $(\omega_{\alpha,t}(z), \omega_{\beta,t}(z))$, and Proposition [A.14] will follow from the exact same proof as that of [S] Proposition 3.1.

As in [S], we define the $t$-dependent domain $D' \equiv D'(\sigma, \tau, \eta_0) := D_{in} \cup D_{out}$ as

$$D_{in} \equiv D_{in}(\tau, \eta_0) := \{ E + i \eta \in \mathbb{C}_+ : E - E^{+}_t, t \in [-\tau, N^{-1+\sigma}], \Im z \in [N^{-1+\sigma}, 0] \},$$

$$D_{out} \equiv D_{out}(\delta, \eta_0) := \{ E + i \eta \in \mathbb{C}_+ : E \in [E^{+}_t + N^{-1+\sigma}, \tau^{-1}], \Im z \in (0, \eta_0) \}.$$

The upper bound for the distance between subordination functions, is established in the following lemma;

**Lemma A.11.** Let $\epsilon > 0$ be fixed and $\tau$ be as in Lemma [A.7]. Then for $N$ sufficiently large, the following hold uniformly over $t \in [0, 1]$ and $z \in D'$ for some $\eta_0$;

$$|\omega_{A,t}(z) - \omega_{\alpha,t}(z)| + |\omega_{B,t}(z) - \omega_{\beta,t}(z)| = O\left(\frac{N^{-1+\epsilon}}{\sqrt{|z - E^{+}_t|}}\right),$$

(A.23)

$$|S_{AB}| \sim \sqrt{|z - E^{+}_t|}, \quad \Im m_{\phi}(z) = O(\sqrt{|z - E^{+}_t|}),$$

and $\Im m_{\phi}(z) \sim \sqrt{|E^{+}_t - z|}$ when $z \in D_{in}$. Furthermore, the following stronger bound hold for $z \in D_{out} \cap \{ z : \Im z \leq N^{-1} \};$

$$|\Im \omega_{A,t}(z) - \Im \omega_{\alpha,t}(z)| + |\Im \omega_{B,t}(z) - \Im \omega_{\beta,t}(z)| = O\left(\frac{(\Im \omega_{\alpha,t} + \Im \omega_{\beta,t})N^{-1+\epsilon} + \Im z}{\sqrt{|z - E^{+}_t|}}\right).$$

(A.24)

The proof of Lemma [A.11] can be further divided into two steps; firstly we prove (A.23) in the regime where $\Im z = \eta$ is large enough, and secondly we prove a stability result that strengthens a given upper bound for the LHS of (A.24) into the RHS of (A.23).

The first step of the proof of Lemma [A.11] is dealt with the following lemma, which corresponds to [S] Lemma A.2. To summarize, it enables us to bound the LHS of (A.23) in terms of $\Phi_{\alpha, \beta}$, defined in (3.3). For later uses we included the corresponding result for $\Phi_{AB}$.

**Lemma A.12.** Let $(r_1, r_2)$ be either $(\alpha, \beta)$ or $(A, B)$ and let $\eta_0 > 0$. For each $t \in [0, 1]$, let $\tilde{\omega}_{1,t}, \tilde{\omega}_{2,t} : \mathbb{C}_{\eta_0} \rightarrow \mathbb{C}_+$ be analytic functions where $\mathbb{C}_{\eta_0} = \{ z \in \mathbb{C}_+ : \Im z \geq \eta_0 \}$. Assume that there is a constant $C > 0$ such that the following hold for all $t \in [0, 1]$ and $z \in \mathbb{C}_{\eta_0}$;

$$|\Im \tilde{\omega}_{1,t}(z) - \Im \omega_{1,t}(z)| \leq C,$$

(A.25)

$$|\Im \tilde{\omega}_{2,t}(z) - \Im \omega_{2,t}(z)| \leq C,$$

(A.26)

$$|\tilde{r}_{1}(z) - r_{1}(z)| \leq C,$$

(A.27)

Then there exists a constant $\eta_0$ with $\eta_0 \geq \eta_0$ such that

$$|\tilde{\omega}_{1}(z) - \omega_{1}(z)| \leq 2|r_{1}(z)|,$$

(A.28)

where $r_{1}(z)$ denotes the two-dimensional vector $(r_{1}(z), r_{1}(z)).$

Since $\tilde{\mu}$ has the same properties as $\mu_0$ except for an additional mass of size $t$, we can easily see that the proof of [S] Lemma A.2 applies to Lemma [A.12]. Details of the proof is left to interested readers.

Next, we establish the local stability result which is used to extend the bound to smaller $\eta$.

**Lemma A.13.** Let $\tau, \eta_0 > 0$ be as in Lemma [A.7]. There exist some constants $c, C > 0$ depending only on $\mu_{\alpha}, \mu_{\beta},$ and $k_0$ such that the following holds for all $t \in [0, 1]$ and $z_0 \in D_t(0, \infty);$ if

$$|\omega_{A,t}(z_0) - \omega_{\alpha,t}(z_0)| + |\omega_{B,t}(z) - \omega_{\beta,t}(z)| \leq c \min(k_0, |S_{\alpha, \beta}(z_0)|)$$

(A.29)

for the constant $k_0$ from Lemma [A.7] then

$$|\omega_{A,t}(z) - \omega_{\alpha,t}(z)| + |\omega_{B,t}(z) - \omega_{\beta,t}(z)| \leq C \frac{|r(z)|}{|S_{\alpha, \beta}(z)|},$$

(A.30)

where $r(z) := (r_{1}(z), r_{2}(z)) := \Phi_{\alpha, \beta}(\omega_{A}(z), \omega_{B}(z), z)$. Furthermore, there exists a constant $C' > 0$ depending only on $\mu_{\alpha}, \mu_{\beta}, k_0$, and $\eta_0$ such that for all $t \in [0, 1]$ and $z_0 \in D_t(0, \eta_0)$, (A.29) implies

$$|r(z)| \leq C'd.$$
Specifically, we take \( c \) and \( C \) so that \( c < 1/3, \ c^{-1} > 4c_1^{-3}c_2(c_1^{-2}c_2 + 1), \) and \( C > 4(c_1^{-2}c_2 + 1) \), where
\[
c_1 := k_0/3, \quad \text{and} \quad c_2 := \mu_{0,0}(R) + \mu_{0,0}(R) + 1.
\]

**Proof of Lemma A.13** For simplicity, we abbreviate
\[
\omega_{\alpha,t}(z_0) = \omega_\alpha, \quad \omega_{A,t}(z_0) = \omega_A, \quad \omega_A - \omega_\alpha = \Delta \omega_1,
\]
\[
\omega_{\beta,t}(z_0) = \omega_\beta, \quad \omega_{B,t}(z_0) = \omega_B, \quad \omega_B - \omega_\beta = \Delta \omega_2.
\]
As in [4 Proposition 4.1], we first consider the Taylor expansion of \( F_{\alpha,t}(\omega_A) \) around \( \omega_\alpha; \)
\[
|F_{\alpha,t}(\omega_A) - F_{\alpha,t}(\omega_\alpha) - F'_{\alpha,t}(\omega_\alpha)\Delta \omega_1| \leq |\Delta \omega_1|^2 \sup_{|\omega - \omega_\alpha| \leq k_0/3} |F''_{\alpha,t}(\omega)| \leq c_1^{-3}c_2|\Delta \omega_1|^2, \tag{A.32}
\]
where we used Lemmas A.2 and A.7 and (A.29) in the last inequality. We have the same bound with \( \alpha, A \) replaced by \( \beta, B. \)

On the other hand by the definition of \( r_1(z) \), we have
\[
F_{\alpha,t}(\omega_A) - F_{\alpha,t}(\omega_\alpha) = \Delta \omega_1 + \Delta \omega_2 + r_1(z_0). \tag{A.33}
\]
Combining (A.32) and (A.33), we find that
\[
|(F'_{\alpha,t}(\omega_\alpha) - 1)\Delta \omega_1 - \Delta \omega_2| \leq c_1^{-3}c_2\|\Delta \omega\|^2 + \|r(z_0)\|, \tag{A.34}
\]
where \( \Delta \omega = (\Delta \omega_1, \Delta \omega_2) \). Now we take a linear combination of (A.32) and its counterpart with switched indices, so that
\[
|S_{\alpha\beta}(z_0)||\Delta \omega_1| \leq (|F'_{\beta,t}(\omega_\beta) - 1| + 1) (c_1^{-3}c_2\|\Delta \omega\|^2 + \|r(z_0)\|). \tag{A.35}
\]
Due to (A.29), we can solve (A.35) as a quadratic inequality for \( \|\Delta \omega\| \), so that
\[
\|\Delta \omega\| \leq 4|S_{\alpha\beta}(z_0)|^{-1}(c_1^{-2}c_2 + 1)|r(z_0)|,
\]
where we used \( |F'_{\beta,t}(\omega_\beta) - 1| \leq c_1^{-2}c_2 \) from Lemmas A.2 and A.7.

Finally, we prove (A.31). We first note that
\[
r_1(z_0) = r_1(z_0) - \Phi_A(\omega_A, \omega_B, z_0) = F_{\alpha,t}(\omega_A) - F_{\alpha,t}(\omega_\alpha).
\]
Then, following [8 (3.10)] we can easily see that \( |m_\alpha(\omega_A) - m_\alpha(\omega_\alpha)| \leq Cd \) from Lemma A.7. Thus by the definition of \( F_{\mu,t} \), the result follows once we have
\[
|m_\alpha(\omega_A)| \geq c \quad \text{and} \quad |m_\alpha(\omega_\alpha)| \geq c
\]
for some constant \( c > 0 \). To see this, we observe from the proof of Lemma A.7 that \( \Re \omega_\alpha - E^+_{\alpha} > k_0/2 \) or \( \Im \omega_\alpha > k_0/2 \) should hold. This implies either one of the following is true;
\[
\Re \omega_\alpha - E^+_{\alpha} > k_0/6 \quad \text{or} \quad \Im \omega_\alpha > k_0/6. \tag{A.36}
\]
Combining (A.36) with Lemma A.8 we get \( |m_\alpha(\omega_A)| \geq c \) and thus \( |m_\alpha(\omega_\alpha)| \geq c \). This concludes the proof of Lemma A.13. \( \square \)

**Proof of Lemma A.11** Large portion of the proof is identical that of [8 Lemma 3.12], and we focus on highlighting the difference rather than explaining the details to avoid repetition.

We first prove (A.23). By [8 Lemma 4.4] and references therein, we find that the subordination functions \( \omega_{\alpha,t} \) and \( \omega_{\beta,t} \) are also Pick functions whose representations satisfy the following;
\[
\omega_{\alpha,t}(z) - z = m_{\mu_{\alpha,t}}(z), \quad \omega_{A,t}(z) - z = m_{\mu_{A,t}}(z), \quad \mu_{\alpha,t}(R) = \mu_{\beta,t}(R), \quad \mu_{A,t}(R) = \mu_{B,t}(R). \tag{A.37}
\]
In particular \( \omega_{\alpha,t}(z) - z \) and \( \omega_{A,t}(z) - z \) are both \( O(\Im z^{-1}) \) uniformly over \( t \in [0, 1] \) and \( \Im z > \eta_0 \) for some \( \eta_0 > 0 \) Thus, taking large enough \( \eta_0 \) and applying Lemma A.12 with the choices \( \tilde{\omega}_{1,t} = \omega_{A,t} \) and \( \tilde{\omega}_{2,t} = \omega_{B,t} \), we obtain
\[
|\omega_{A,t}(z) - \omega_{\alpha,t}(z)| \leq 2|z|, \quad \Re z \in [E^+_{\alpha} - \tau, \tau^{-1}], \quad \Im z = \eta_0.
\]

35
Furthermore, replicating the proof of (A.31) yields \( \|r(z)\| \leq C_{\eta_0} \mathbf{d} \) for some constant \( C_{\eta_0} \) depending on \( \eta_0 \).

Now we take \( N \) to be sufficiently large so that
\[
4C_{\eta_0} \mathbf{d} \leq c \min(\eta_0, \inf\{ |S_{\alpha\beta}(E + i\eta_0)| : E \in [E_{i,t} - \tau, \tau^{-1}] \})
\]
where \( c \) is the constant in Lemma (A.13) Here we used the fact that
\[
|S_{\alpha\beta}(E + i\eta_0)| \geq 1 - \eta_0^{-4} \tilde{\mu}_{\alpha,1}(\mathbb{R}) \tilde{\mu}_{\beta,1}(\mathbb{R}),
\]
which follows from (A.11) and the definition of \( S_{\alpha\beta} \). Thus by Lemma (A.13) we have
\[
|\omega_{A,t}(z) - \omega_{A,t}(z)| + |\omega_{B,t}(z) - \omega_{B,t}(z)| \leq C_* \frac{d}{|S_{\alpha\beta}(z)|}, \quad z = E + i\eta_0 \in \mathcal{D}_r(0, \infty).
\]
(A.38)
where \( C_* = CC' \) and \( C, C' \) are the constants in Lemma (A.13) applied to the domain \( \mathcal{D}_r(0, \eta_0) \).

Following [3] Lemma 3.12, we take \( \eta_m \) to be the smallest number for which (A.38) holds for all \( z = E + i\eta \) with \( E \in [E_{i,t} - \tau, \tau^{-1}] \) and \( \eta \in [\eta_0, \eta_0] \). By above, such \( \eta_m \) must exist and we have \( \eta_m \leq \eta_0 \).

Suppose on the contrary that \( \eta_m > N^{-1+\sigma} \), and take \( \eta_m' = \eta_m - N^{-2} \) and \( E \in [E_{i,t} - \tau, \tau^{-1}] \). By (A.37), we find that
\[
|\omega_{A,t}(E + i\eta_m) - \omega_{A,t}(E + i\eta_m')| \leq C \eta_m^{-2}(\eta_m - \eta_m')^2 \leq C N^{2\sigma - 2}
\]
for some numeric constant \( C > 0 \), and that the same bound holds for \( \omega_{\beta,t}, \omega_{A,t} \), and \( \omega_{B,t} \). Then we have
\[
|\omega_{A,t}(E + i\eta_m) - \omega_{A,t}(E + i\eta_m')| \leq C \frac{d}{|S_{\alpha\beta}(E + i\eta_m)|} + C N^{2\sigma - 2} \leq C (N^{-4/2} d + N^{2\sigma - 2}),
\]
where we used the fact that \( |S_{\alpha\beta}(z)| \sim \sqrt{\kappa + \eta} \). Again using the asymptotics for \( |S_{\alpha\beta}| \), we have
\[
N^{-\sigma/2} d + N^{2\sigma - 2} \leq N^{(-1+\gamma)/2} \leq |S_{\alpha\beta}(E + i\eta_m')|,
\]
where we used Assumption (2.3). Taking large enough \( N \), we see that the point \( E + i\eta_m \) satisfies the assumptions of Lemma (A.13) so that (A.38) holds true at the point. Since \( E \) was arbitrary chosen in \( [E_{i,t} - \tau, \tau^{-1}] \), we obtain a contradiction to \( \eta_m > N^{-1+\sigma} \). Therefore we conclude that (A.38) holds for all \( z \in \mathcal{D}_r(N^{-1+\sigma}, \eta_0) \), and (A.23) for all \( z \in \mathcal{D}_0 \). The proof for \( \mathcal{D}_\text{out} \) is exactly the same, except we take \( E \in [E_{i,t} + N^{-1+\sigma}], \eta_m \in (0, N^{-1+\sigma}] \), and \( \eta_m' = \eta_m \).

The rest of the proof is exactly the same as [3] Lemma 3.13 and we omit the details for simplicity.

\[\Box\]

**Proof of Proposition (A.1)** The proof is almost identical to that of [3] Proposition 3.1, and the only difference is that the \( F \)-transform therein should be replaced by \( F_2 \) defined in (3.1). In particular, using Lemmas (A.7) and (A.11) in place of Lemmas 3.7 and 3.12 of [3], one can easily check that the proof of [3] Proposition 3.1] applies verbatim. We omit further details.

**B Local laws for \( H_t \)**

In the rest of the appendix, we prove local laws for \( H_t \) following the same strategy as in [3]. The proofs in [3] have to be carefully modified in order to deal with the effect of DBM. Due to similarity, we mainly focus on explaining such modification and refer to [3] whenever the same calculation therein applies.

Recall the definition of \( \mathcal{D}_r(a, b) \) from (3.1). The precise statement of the local law is as follows;

**Theorem B.1 (Local laws for \( H_t \)).** Suppose that Assumption (2.3) holds. Let \( \tau > 0 \) be a sufficiently small constant, \( \sigma > 0 \), and \( d_1, \ldots, d_N \in \mathbb{C} \) be deterministic complex numbers satisfying \( \max_{i \in [1,N]} |d_i| \leq 1 \). Then we have
\[
\left| \frac{1}{N} \sum_t d_t \left( G_{ii} - \frac{1}{a_i - \omega_A(z)} \right) \right| \prec \frac{1}{N \eta} \quad \forall t \in [0,1]
\]
uniformly over \( t \in [0,1] \) and \( z \in \mathcal{D}_r(N^{-1+\sigma},1) \). Furthermore, we have
\[
\max_{i,j \in [1,N]} |G_{ij}(z) - \delta_{ij} \frac{1}{a_i - \omega_A(z)}| + |(U^* G)_{ij}| \leq \left( \frac{\max |m_{ij}(z)|}{N \eta} + \frac{1}{N \eta} \right)
\]
uniformly over the same domain for \( t \) and \( z \). The same results hold true if we replace \( G, a_i, \) and \( \omega_A(z) \) by \( \mathcal{G}, b_i, \) and \( \omega_B(z) \), respectively.
We postpone the proof of Theorem B.1 and first prove Proposition 4.4 with the aid of Theorem B.1.

**Proof of Proposition 4.4.** Recall that the spectral parameter $z$ in Proposition 4.3 is restricted to $\{E+i\gamma_0 : E \in [E_1, E_2]\}$. To prove (B.3), we write

\[
|\gamma t m(z) - m_{\mu_1}(E_+, t)| \leq \frac{1}{N} \sum_i \left| G_{ii}(z/\gamma) - \frac{1}{a_i - \omega_A(z/\gamma)} \right| + |m_A(\omega_A(z/\gamma)) - m_A(\omega(z/\gamma))| + |m_A(\omega(z/\gamma)) - m_{\mu_1}(E_+))|.
\]

For the first term on the RHS of (B.3), the averaged local law (B.1) directly implies the bound $O(N^{-1/3+\epsilon})$. Combining Lemmas A.7 and A.11 proves that the second term $O(N^{-1/3+\epsilon})$. Finally for the last term, we apply Lemma A.7, Proposition A.9, and $O$ calculations in [8] for the first term and [24] for the second, we see that estimating the rightmost side of (B.2) is $O(N^{-1/3+\epsilon})$ for our choice of $z$, we have the bound. For the last estimate (B.5), we refer to (4.3) below.

As in Theorem B.1, we always work with the free convolution $\tilde{\mu}_t = \mu_A \boxplus \mu_B \boxplus \mu_{sc}^{(t)}$ rather than its limit $\mu_t$. Thus we use the shorthand notation $m_0(z) = m_{\tilde{\mu}_t}(z)$ without any confusion. In the following sections, we abbreviate $D \equiv D_1(N^{-1+\sigma}, 1)$.

**B.1 Outline of the proof of local law**

To simplify the presentation, we introduce the following control parameters depending on $N$, $t$, and $z$.

\[
\Psi \equiv \Psi(z) := \sqrt{\frac{1}{N\eta}}, \quad \Pi \equiv \Pi(z) := \sqrt{\frac{\text{Im} m_{\mu_1}}{N\eta}},
\]

\[
\Pi_i \equiv \Pi_i(z) := \sqrt{\frac{\text{Im}(G_{ii} + G_{ii})}{N\eta}}, \quad \Pi_i \equiv \Pi_i(z) := \sqrt{\frac{\text{Im}(WG)_{ii}}{N\eta}}.
\]

Before proceeding to the actual proof, we first present the outline of the proof of Theorem B.1. We first define random functions $\omega_A$ and $\omega_B$ in $z$ as follows;

\[
\omega_A := z - \frac{\text{tr} \tilde{B} G}{\text{tr} G} + t \text{tr} G, \quad \omega_B := z - \frac{\text{tr} A G}{\text{tr} G} + t \text{tr} G.
\]

These functions will serve as random approximates of the genuine subordination functions $\omega_A$ and $\omega_B$ associated to the free convolution $\mu_A \boxplus \mu_B \boxplus \mu_{sc}^{(t)}$. Note that these functions have an additional term $t \text{tr} G$ compared to those used in [8]. With these choices, we follow the same three-step-strategy as in [8].

The first step is to prove the entrywise subordination, that is, estimates of the form

\[
\left| G_{ij} - \delta_{ij} \frac{1}{a_i - \omega_A^c} \right| \prec \Psi.
\]

More specifically, we write

\[
zG_{ij} + \delta_{ij} - a_i G_{ij} = (\tilde{B} G)_{ij} + \sqrt{t} (W G)_{ij}
\]

using $zG + I = HG$ and apply Gaussian integration by parts to the rightmost side of (B.6). Following calculations in [8] for the first term and [24] for the second, we see that estimating the rightmost side of (B.6) is equivalent to upper bounds for $Q_{ij}$ and $L_{ij}$ defined as

\[
Q_{ij} := (\tilde{B} G)_{ij} \text{tr} G - G_{ij} \text{tr}(\tilde{B} G) \quad \text{and} \quad L_{ij} := (W G)_{ij} + \sqrt{t} \text{tr}(G)G_{ij}.
\]

In fact, after some algebraic calculation, we arrive at

\[
G_{ij} - \delta_{ij} \frac{1}{a_i - \omega_A^c} = -\frac{Q_{ij}}{(a_i - \omega_A^c) \text{tr} G} - \frac{\sqrt{t} L_{ij}}{a_i - \omega_A^c}.
\]
In order to apply Stein’s lemma to $Q_{ij}$, we use partial randomness decomposition in Section 3.1 to extract Gaussian random variables from the Haar unitary matrix $U$ as in [8]. Consequently, the quantity $Q_{ij}$ can be controlled by the two quantities in (4.10) of [8], namely

$$S_{ij} := h_i^* \tilde{B}^{(i)} G e_j, \quad T_{ij} := h_i^* G e_j,$$

where $h_i$ and $\tilde{B}^{(i)}$ are defined in Section 3.1. Indeed, we can see from discussions below (4.10) of [8] that quantities in (B.8) are directly connected to $(\tilde{B}G)_{ij}$. Following [8], we find that it easier to work with auxiliary quantities $P_{ij}$ and $K_{ij}$ instead of $Q_{ij}$, defined by

$$P_{ij} := (\tilde{B}G)_{ij} trG - G_{ij} tr(\tilde{B}G) + (G_{ij} + T_{ij}) \Upsilon,$$

$$K_{ij} := T_{ij} + (b_i T_{ij} (\tilde{B}G)_{ij}) trG - (G_{ij} + T_{ij}) tr(\tilde{B}G),$$

where

$$\Upsilon := tr(\tilde{B}G) - (tr(\tilde{B}G)^2 + trG tr(\tilde{B}G)).$$

On the other hand, calculations for $L_{ij}$ involve another quantity $J_i$ defined as

$$J_i := (W G W)_{ii} - trG + \sqrt{t} tr(G(W G)_{ii}).$$

In summary, the first step mainly concerns estimates for the four quantities $J_i, L_{ij}, P_{ij},$ and $K_{ij}$. The corresponding result is Proposition [13.5], which is proved in the next subsection.

In the second step, we estimate the distance between $\omega^*\Lambda$ and $\omega\Lambda$. The bulk of the proof is devoted to rough and optimal fluctuation averaging results for $Q_{ii}$ and $L_{ii}$. In the rough fluctuation averaging, Propositions 6.1, 6.2, we prove that

$$\left| \frac{1}{N} \sum_i d_i Q_{ii} \right| \prec \Psi \Pi \quad \text{and} \quad \left| \frac{1}{N} \sum_i d_i L_{ii} \right| \prec \Pi^2,$$

(B.9)

for generic bounded weights $d_1, \ldots, d_N$. Note that the bounds in (B.9) are much smaller than the bound $\Psi$ in (3.2) due to an averaging effect of fluctuations. And then in the optimal fluctuation averaging, Proposition 6.3, we take a specific weights for $Q_{ii}$ and $L_{ii}$ and consider a specific combination of two averages. Our choice leads to an improved bound $\Pi^2$ for the first term in (B.9) and to an estimate of the form

$$|S_{AB}(z) \Lambda_A(z)^2 + T_A(z) \Lambda_A(z) + O(\Lambda_A(z)^3)| \prec \Pi^2,$$

(B.10)

where $\Lambda_A(z) = \omega^*\Lambda(z) - \omega\Lambda(z)$. The bound (B.10) eventually results in the bound $|\Lambda_A(z)| \prec \Psi^2$. Details for the second step can be found in Section C.

Throughout both the first and second steps, we fix a spectral parameter $z$, assume that a weak, probabilistic bound holds at the point $z$, and use this assumption as an input. In the third and final step, we prove a weak local law to ensure that this a priori bound is in fact true in the whole domain. More specifically, we invoke the proofs in previous steps to prove weaker but quantitative versions of entrywise subordination and fluctuation averaging, in the sense that they do not depend on a probabilistic input. Then we use a bootstrapping argument to conclude a weak local law, Theorem D.1. Feeding the weak law back to the first and second steps and using another bootstrapping argument lead to the final result. This step is presented in Section D.

### B.2 Entrywise subordination

We introduce the following notations for the errors we need to control:

$$\Lambda_{ij} := \left| G_{ij} - \frac{\delta_{ij}}{a_i - \omega A} \right|, \quad \Lambda_{ij}^\prime := \left| G_{ij} - \frac{\delta_{ij}}{a_i - \omega^*\Lambda} \right|, \quad \Lambda_L := \max_{i,j} |L_{ij}|,$$

$$\Lambda_c := \max \Lambda_{ij}, \quad \Lambda_c^\prime := \max \Lambda_{ij}^\prime, \quad \Lambda_T := \max_{i,j} |T_{ij}|.$$

We also write $\tilde{\Lambda}_{ij}, \tilde{\Lambda}_{ij}, \tilde{\Lambda}_T, \tilde{\Lambda}_c, \tilde{\Lambda}_c^\prime$ to represent their analogues obtained by switching the roles of $(A, B)$, $(U, U^*)$, and $(W, W)$. For example, we write

$$\tilde{\Lambda}_{ij} := \left| g_{ij} - \frac{\delta_{ij}}{b_i - \omega B} \right|.$$
Finally, we take a collection of smooth cut-off function \( \varphi_L \equiv \varphi : \mathbb{R} \to \mathbb{R} \) indexed by \( L > 0 \) such that

\[
\varphi(x) = \begin{cases} 
1 & |x| \leq L, \\
0 & |x| \geq 2L, 
\end{cases} \quad \sup_{x \in \mathbb{R}} |\varphi'(x)| \leq CL^{-1}. \tag{B.12}
\]

We introduce another notation that strengthens the notion of stochastic dominance. For an \( N \)-dependent random variable \( X \) that may also depend on \( t \) and \( z \), we write \( X = O_\prec(Y) \) for a positive deterministic function \( Y \) of \( N, t, z \) when the following holds: For any fixed \( \epsilon > 0 \) and \( p \in \mathbb{N} \), there exists an \( N_0 \in \mathbb{N} \) depending only on \( \epsilon \) and \( p \) such that

\[
\mathbb{E}[|X|^p] \leq N^\epsilon Y^p
\]

whenever \( N \geq N_0 \). In this case, we often write \( O_\prec(Y) \) in place of \( X \). Indeed, we can easily see that \( X = O_\prec(Y) \) implies \( X \prec Y \) by Markov’s inequality.

**Lemma B.2.** Let \( \mathcal{L} > 0 \) and \( p \in \mathbb{N} \) be fixed and define

\[
\Gamma_{i0} := |G_{ii}|^2 + |(GW)_{ii}|^2 + |(WG)_{ii}|^2 + |\text{tr} G|^2
\]

Then there exist \( N_0, C_{p, L} > 0 \) depending only on \( \mathcal{L} \) and \( p \) such that

\[
\mathbb{E}[|I_i(\Gamma_{i0})|^2] \leq C_{p, L} \Psi_{2p}
\]

for all \( N \geq N_0, i \in [1, N], t \in [0, 1] \) and \( z \in D \).

**Proof.** We use the following shorthand notations throughout the proof:

\[
i_i^{(p, q)} = J_i^p \Gamma_i^q \varphi(\Gamma_{i0})^{p+q} \quad p, q \in \mathbb{N}, \tag{B.13}
\]

with conventions \( i_i^{(0, 0)} = 1 \) and \( i_i^{(-1, 1)} = 0 \). Using this notation, we write

\[
\mathbb{E}[I_i^{(p, q)}] = \mathbb{E}[(GW)_{ii} \varphi(\Gamma_{i0})]^{(p-1, p)} - \mathbb{E}[	ext{tr} G \varphi(\Gamma_{i0})]^{(p-1, p)} + \sqrt{\mathbb{E}[	ext{tr} G(WG)_{ii} \varphi(\Gamma_{i0})]^{(p-1, p)}}. \tag{B.14}
\]

Applying Stein’s lemma to the first term yields

\[
\mathbb{E}[(GW)_{ii} \varphi(\Gamma_{i0})]^{(p-1, p)} = \sum_k \mathbb{E}[(WG)_{ik} W_{ki} \varphi(\Gamma_{i0})]^{(p-1, p)} \]

\[
= \mathbb{E}[	ext{tr} G \varphi(\Gamma_{i0})]^{(p-1, p)} - \sqrt{\mathbb{E}[	ext{tr} G(WG)_{ii} \varphi(\Gamma_{i0})]^{(p-1, p)}} + \frac{2p-1}{N} \sum_k \mathbb{E}[(WG)_{ik} \frac{\partial J_i}{\partial W_{ik}} \varphi(\Gamma_{i0})]^{(p-2, p)} + \frac{p-1}{N} \sum_k \mathbb{E}[(WG)_{ik} \frac{\partial J_i}{\partial W_{ik}} \varphi(\Gamma_{i0})]^{(p-1, p-1)} \tag{B.15}
\]

The first two terms of \( \text{(B.15)} \) are canceled with the last two terms of \( \text{(B.14)} \). For the remaining terms, we use the following lemma;

**Lemma B.3.** Let \( X_i \) be one of \( G_{ii}, (GW)_{ii}, (WG)_{ii} \) or \( (WG)_{ii} \). Then there exists a constant \( C_L \) depending only on \( \mathcal{L} \) such that the following hold for all \( i \in [1, N], t \in [0, 1] \), and \( z \in D \):

\[
\frac{1}{N} \sum_k (WG)_{ik} \frac{\partial X_i}{\partial W_{ik}} \varphi(\Gamma_{i0}) \leq C_L (N^{-1} + \Pi_i^2 + (\Pi_i^W)^2) \varphi(\Gamma_{i0}), \tag{B.16}
\]

\[
\frac{1}{N} \sum_k (WG)_{ik} \frac{\partial \text{tr} G}{\partial W_{ik}} \varphi(\Gamma_{i0}) \leq C_L \Psi^2 (N^{-1} + \Pi_i^2 + (\Pi_i^W)^2) \varphi(\Gamma_{i0}). \tag{B.17}
\]

Furthermore, the same bounds hold true if we replace \( X_i \) or \( \text{tr} G \) with its complex conjugate or if \( \Gamma_{i0} \) is replaced by a larger quantity than \( \Gamma_{i0} \).
Proof of Lemma B.3 Due to similarity, we only consider the first term of (B.16) with the choice $X_i = (WG_W)_{ii}$. Computing the derivative explicitly, we find that
\[ \sum_k (WG_k)_{ik} \frac{\partial (WG_W)_{ik}}{\partial W_{ik}} = (WG^2 W)_{ii} - \sqrt{t} (WG)_{ii} (WG^2 W)_{ii} + (WG)_{ii}^3. \]
Applying Cauchy-Schwarz inequality to the entry $(WG)_{ii}$ gives
\[ \| (WG^2 W)_{ii} \| \leq \| GW e_i \| \| G^* W e_i \| \leq \frac{\text{Im}(WG)_{ii}}{\eta} = N (\Pi^W)_{ii}, \]
and the definition of $\Gamma_{ii}$ yields $| (WG)_{ii} | \phi(\Gamma_{ii}) \leq \sqrt{2L} \phi(\Gamma_{ii})$. Thus we conclude
\[ \frac{1}{N} \sum_k (WG_k)_{ik} \frac{\partial (WG_W)_{ik}}{\partial W_{ik}} \phi(\Gamma_{ii}) \leq C L (N^{-1} + (\Pi^W)_{ii}^2) \phi(\Gamma_{ii}) \]
as desired.

After several applications of Leibniz and chain rules, we see that the coefficient of $j_i$ in each of the third, fourth, and fifth terms of (B.15) can be further decomposed into quantities in Lemma B.3 up to a factor of $C_L$. On the other hand, we also have
\[ \Pi^W_{ii} \phi(\Gamma_{ii}) \leq C_L \Psi^2 \phi(\Gamma_{ii}) \text{ and } (\Pi^W_{ii})^2 \phi(\Gamma_{ii}) \leq \Psi^2 (|J_i| + C_L) \phi(\Gamma_{ii}), \]
where we used the definition of $J_i$ in the second inequality. Plugging these inequalities into (B.14), we find that
\[ \mathbb{E}_{\eta}^{|j_i|^{(p-p)}} \leq p C_L (\mathbb{E}[\Psi^2 |J_i| \phi(\Gamma_{ii})]^{2p-1} + p C_L (\mathbb{E}[\Psi^2 |J_i| \phi(\Gamma_{ii})]^{2p-2} + p C_L (\mathbb{E}[\Psi^2 |J_i| \phi(\Gamma_{ii})]^{2p}). \]
Applying Jensen’s inequality to the first two terms of (B.15), we obtain that $x = \Psi^{-1} \mathbb{E}[|j_i|^{(p-p)}]^{1/2p}$ satisfies the quadratic inequality
\[ x^2 \leq C_{p,L}(\Psi x + 1), \]
so that $x \leq C_{p,L}$ for a constant $C_{p,L}$ depending only on $p$ and $L$. This concludes the proof of Lemma B.3.

Lemma B.4. Let $\mathcal{L} > 0$ and $p \in \mathbb{N}$ be fixed and define
\[ \Gamma_{ij0} := |(\text{tr} G)|^2 + |G_{ii}|^2 + |G_{jj}|^2 + |G_{ij}|^2 + |(WG)_{ii}|^2 + |(GW)_{ii}|^2, \quad (B.19) \]
Then there exist $N_0 \in \mathbb{N}$ and $C_{p,L} > 0$ depending on $\mathcal{L}$ and $p$ such that
\[ \mathbb{E}[|L_{ij} \phi(\Gamma_{ij0})|^{2p}] \leq C_{p,L} \Psi^{2p} \quad (B.20) \]
for all $N \geq N_0$, $i,j \in [1, N]$, $t \in [0, 1]$, and $z \in \mathcal{D}$.

Proof. We take $I_{ij}^{(p-q)} = L_{ij}^{(p-q)} \phi(\Gamma_{ij})^{p+q}$ for $p, q \in \mathbb{N}$, so that the LHS of (B.20) is equal to $\mathbb{E}[|I_{ij}^{(p-q)}|]$. Following lines of proof of Lemma B.2 we can prove that there are random variables $k_1, k_2$, and $k_3$ with
\[ \mathbb{E}[|I_{ij}^{(p-q)}|] = \mathbb{E}[k_1 I_{ij}^{(p-1,q-1)}] + \mathbb{E}[k_2 I_{ij}^{(p-2,q)}] + \mathbb{E}[k_3 I_{ij}^{(p-1,q-1)}] \quad (B.21) \]
such that
\[ |k_1| \phi(\Gamma_{ij0}) \leq C_{L} (N^{-1} + \Pi^2 + \Pi^2 + (\Pi^W)^2) \phi(\Gamma_{ij0}), \quad i = 1, 2, 3. \]
The proof of this fact follows from the following off-diagonal variant of Lemma B.3. We omit its proof since it is identical to that of Lemma B.3 except we use $\| G e_i \|^2 \leq N \Pi^2$ as an additional input.

Lemma B.5. Let $X_{ij}$ be one of $G_{ij}$, $(WG)_{ij}$, $(GW)_{ij}$, $G_{jj}$, $G_{ij}$, $G_{ji}$. Then there exists a constant $C_L$ depending only on $\mathcal{L}$ such that the following hold for all $i,j \in [1, N]$, $t \in [0, 1]$, and $z \in \mathcal{D}$:
\[ \frac{1}{N} \sum_k G_{ik} \frac{\partial X_{ij}}{\partial W_{ki}} \phi(\Gamma_{ij0}) \leq C_{L} (N^{-1} + \Pi^2 + \Pi^2 + (\Pi^W)^2) \phi(\Gamma_{ij0}), \]
\[ \frac{1}{N} \sum_k G_{ik} \frac{\partial \text{tr} G}{\partial W_{ki}} \phi(\Gamma_{ij0}) \leq C_{L} \Psi^2 (N^{-1} + \Pi^2 + \Pi^2 + (\Pi^W)^2) \phi(\Gamma_{ij0}). \]
The same set of inequalities holds true if $X_{ij}$ and $\text{tr} G$ are replaced by their complex conjugates.
We now deduce (B.20) from (B.21). As in Lemma 3.2, we further bound the control parameters using
\[ (\Pi_1^t + \Pi_1^t)^2 \varphi(\Gamma_{ij}) \leq C_{\varphi} \Psi^2 \quad \text{and} \quad (\Pi_1^W)^2 \varphi(\Gamma_{ij}) \leq C_{\varphi} \Psi^2 (|J_t| + L)^2 \varphi(\Gamma_{ij}). \] (B.22)
Plugging (B.22) into (B.21) and using Jensen and Hölder inequalities, we obtain
\[ \mathbb{E}[t_{ij}^{(p,p)}] \leq C_{\varphi} \Psi^2 \left( (1 + \mathbb{E}|t_{ij}^{(p,p)}|^{1/2p}) \mathbb{E}[t_{ij}^{(p,p)}]^{2p-1} + (1 + \mathbb{E}|t_{ij}^{(p,p)}|^{1/2p}) \mathbb{E}[t_{ij}^{(p,p)}]^{2p-1} \right), \]
where \( j_i \) is defined in (B.13). Since \( \varphi(\Gamma_{ij0}) \leq \varphi(\Gamma_{ij0}) \), Lemma 3.2 implies that
\[ \mathbb{E}[t_{ij}^{(k,k)|2k}] \leq C_{k,k} \Psi, \quad k \in \mathbb{N}. \]
Then we follow the exact same argument as in Lemma 3.2 to conclude (B.20).

Furthermore, we have the similar estimates for \( F_{ij} \) and \( K_{ij} \).

**Lemma B.6.** Let \( \epsilon, L > 0 \) and \( p \in \mathbb{N} \) be fixed and define
\[ \Gamma_{ij} := \Gamma_{ij} + |G_{ij}|^2 + |T_{ij}|^2 + |T_{ij}|^2 + |\text{tr} G|^2 + |\text{tr} B G|^2 + |\text{tr} \tilde{B} G|^2. \] (B.23)

Then there exists \( N_0 \equiv N_0(\epsilon, L, p) \in \mathbb{N} \) such that
\[ \mathbb{E}[|P_{ij} \varphi(\Gamma_{ij})|^{2p}] \leq \Psi^{2p} N^s, \] (B.24)
\[ \mathbb{E}[|K_{ij} \varphi(\Gamma_{ij})|^{2p}] \leq \Psi^{2p} N^s, \] (B.25)
for all \( N \geq N_0 \), \( i, j \in [1, N] \), \( t \in [0, 1] \), and \( z \in \mathcal{D} \).

**Proof.** The proof is similar to that of Lemma 8.3 in [8], and the difference originates from the identity
\[ \tilde{B} G = z G - A G - \sqrt{t W G} + I. \] (B.26)
Due to the additional term \( \sqrt{t W G} \) in the above identity, several new terms arise that do not appear in [8]. We notice that from the definition of \( \Gamma_{ij} \), \( \varphi(\Gamma_{ij0}) \) and \( \varphi(\Gamma_{ij}) \) are positive if \( \varphi(\Gamma_{ij}) \) is. Hence we can apply the estimates from Lemma 3.2 and 3.3.

We first introduce the counterpart of Lemma 5.3 in [8] that handles errors arising along the proof.

**Lemma B.7** (Lemma 5.3 in [8]). Suppose the assumptions of Proposition 5.2 hold. Let \( Q \in M_N(\mathbb{C}) \) be a generic matrix and set \( X_t = I \) or \( B^{(i)} \) and \( X = I \) or \( A \). Then the following hold true for all \( N \in \mathbb{N}, \ t \in [0, 1] \), and \( z \in \mathcal{D} \):

\[
\frac{1}{N} \sum_{k} \left| \frac{\partial}{\partial g_{ik}} e_k^* X_t G e_j \varphi(\Gamma_{ij}) \right| = O_{\omega}((N^{-1})||Q||\varphi(\Gamma_{ij})),
\]
\[
\frac{1}{N} \sum_{k} (e_k^* X_t G) e_j \varphi(\Gamma_{ij}) \leq C_{\omega} \Psi (\Pi^t_1 + \Pi^t_1) \varphi(\Gamma_{ij}),
\]
\[
\frac{1}{N} \sum_{k} \left| \frac{\partial T_{ij}}{\partial g_{ik}} e_k^* X_t G e_j \varphi(\Gamma_{ij}) \right| \leq C_{\omega} \psi (\Pi^t_1 + \Pi^t_1) \varphi(\Gamma_{ij}),
\]
\[
\frac{1}{N} \sum_{k} \left| \frac{\partial}{\partial g_{ik}} e_k^* X_t G e_j \varphi(\Gamma_{ij}) \right| \leq C_{\omega} \psi (\Pi^t_1 + \Pi^t_1 + t(\Pi^W_t)^2) \varphi(\Gamma_{ij}),
\]
\[
\frac{1}{N} \sum_{k} \left| \frac{\partial T_{ij}}{\partial g_{ik}} e_k^* X_t G e_j \varphi(\Gamma_{ij}) \right| \leq C_{\omega} \psi (\Pi^t_1 + \Pi^t_1 + t(\Pi^W_t)^2) \varphi(\Gamma_{ij}),
\]
\[
\frac{\sqrt{t}}{N} \sum_{k} \left| \frac{\partial}{\partial g_{ik}} e_k^* X_t G e_j \varphi(\Gamma_{ij}) \right| \leq C_{\omega} \psi (\Pi^t_1 + \Pi^t_1 + (\Pi^W_t)^2) \varphi(\Gamma_{ij}),
\]
for some constant \( C_{\omega} \) depending only on \( L \). In addition, the same estimates hold if we replace \( \frac{\partial}{\partial g_{ik}} \) and \( \frac{\partial T_{ij}}{\partial g_{ik}} \) by their complex conjugates \( \frac{\partial}{\partial \bar{g}_{ik}} \) and \( \frac{\partial}{\partial \bar{g}_{ik}} \).
Proof. The proof is a straightforward modification of Lemma 5.3 in [S]. The only difference is we use

$$
\|G\tilde{B}_{ij}\| \leq 1 + |z|\|G e_i\| + |a_i|\|G e_i\| + \sqrt{t}\|GW e_i\| \leq 1 + C \sqrt{\frac{\text{Im} G_{ii}}{\eta}} + \sqrt{\frac{t \text{Im}(W G W)_{ii}}{\eta}}.
$$

(B.27)

Note that the last term on the RHS of (B.27) does not appear in [S]. We omit further details. □

With this lemma, we can follow the proof of Lemma 5.3 in [S] verbatim to obtain the recursive moment estimates for \(P_{ij}\) and \(K_{ij}\): Define for \(p, q \in \mathbb{N}\)

$$
m_{ij}^{(p,q)} := P_{ij}^p \tilde{R}_{ij} \varphi(\Gamma_{ij})^{k+1}, \quad n_{ij}^{(p,q)} := K_{ij}^p \tilde{R}_{ij} \varphi(\Gamma_{ij})^{k+1}.
$$

(B.28)

Then we have

$$
\mathbb{E}[m_{ij}^{(p,p)}] = \mathbb{E}[k_1 m_{ij}^{(p-1,p)}] + \mathbb{E}[k_2 m_{ij}^{(p-2,p)}] + \mathbb{E}[k_3 m_{ij}^{(p-3,p)}],
$$

(B.29)

$$
\mathbb{E}[n_{ij}^{(p,p)}] = \mathbb{E}[k_1 n_{ij}^{(p-1,p)}] + \mathbb{E}[k_2 n_{ij}^{(p-2,p)}] + \mathbb{E}[k_3 n_{ij}^{(p-3,p)}].
$$

(B.30)

where \(k_1, k_2, k_3, k'_1, k'_2, k'_3\) are some random variables satisfying

$$
k_1 = O_{\prec}(\Psi), \quad k_2 = O_{\prec}(\Psi^2), \quad k'_2 = O_{\prec}(\Psi^2), \quad k_3 = O_{\prec}(\Psi^2), \quad k'_3 = O_{\prec}(\Psi^2).
$$

(B.31)

As in Lemma B.32, applying Young’s inequality to (B.29) gives the result. □

Now we are ready to present the entrywise subordination;

**Proposition B.8.** Fix \(z \in \mathcal{D}\) and assume that

$$
\Lambda_\sigma(z) \prec N^{-\sigma/4}, \quad \Lambda_L(z) \prec N^{-\sigma/4}, \quad \Lambda_T(z) \prec 1,
$$

(B.32)

and the same set of bounds hold true for \(\tilde{\Lambda}_\sigma, \tilde{\Lambda}_L,\) and \(\tilde{\Lambda}_T\). Then we have for all \(i, j \in [1, N]\) that

$$
|L_{ij}| \prec \Psi(z), \quad |P_{ij}| \prec \Psi(z), \quad \|K_{ij}\| \prec \Psi(z), \quad \text{and} \quad |T(z)| \prec \Psi(z).
$$

(B.33)

Furthermore, we have

$$
\Lambda_\sigma^e \prec \Psi(z) \quad \text{and} \quad \Lambda_T \prec \Psi(z).
$$

(B.34)

The same statements remain true if we switch the roles of \((A, B), (U, U^*),\) and \((W, W)\).

**Proof.** First of all, we remark that the assumption (B.32) implies

$$
\text{tr}(G) = \tilde{m} + O_{\prec}(N^{-\sigma/4}), \\
\text{tr}(\tilde{B}G) = 1 + \omega_B(z)\tilde{m} + O_{\prec}(N^{-\sigma/4}), \\
\text{tr}(\tilde{B}G \tilde{B}) = \omega_B(z)(1 + \omega_B(z)\tilde{m}) + O_{\prec}(N^{-\sigma/4}).
$$

(B.35)

They can be proved in exactly the same way as in (5.20) of [S]. Also the first two estimates in (B.35) combined with the definition of \(\omega_\Lambda^e\) imply

$$
\omega_\Lambda^e(z) = F_{\tilde{m}}(z) + z - \omega_B(z) + O_{\prec}(N^{-\sigma/4}) = \omega_A(z) + O_{\prec}(N^{-\sigma/4}).
$$

(B.36)

The first three estimates in (B.33) are direct consequences of Lemmas B.3 and B.6 since \(\varphi(\Gamma_{ij}) = 1 = \varphi(\Gamma_{ij}^0)\) under the assumption (B.32). The remaining estimates can be proved using the same argument as in Proposition 5.1 in [S]. □
C Fluctuation averaging estimates

C.1 Rough fluctuation averaging for general linear combinations

This section is a counterpart of Section 6 in [8] which gives a rough fluctuation averaging estimate for \( Q_{aa} \) and \( L_a \), Proposition C.1 below. To deal with the contribution of \( W \), we prove a fluctuation averaging estimates for \( L_{aa} \) and \( J_a \). Then we follow the same method as Proposition 6.1 of [8], where the results for \( L_{aa} \) and \( J_a \) are used as additional inputs.

Before proceeding to the proof, we observe that the average of \( \Pi_{aa}^W \) is dominated by \( \Pi \). To be precise, since \( \|W\| < 1 \), we have

\[
\frac{1}{N^2} \sum_a (\Pi_{aa}^W)^2 < t \Pi^2.
\]

Next, we show the fluctuation averaging estimates for \( L_{aa} \) and \( J_a \).

**Proposition C.1.** Fix a \( z \in D \). Suppose the assumptions of Proposition C.1 hold. Let \( d_1, \ldots, d_N \in \mathbb{C} \) be possibly \( H \)-dependent quantities satisfying \( \max |d_i| < 1 \). Assume that for all \( i, j \in [1, N] \),

\[
\frac{\sqrt{7} N}{N^2} \sum_k \frac{\partial d_j}{\partial W_{ki}} e_k Ge_j = O(\Psi^2 \Pi^2),
\]

and the same bounds hold when the \( d_j \)'s are replaced by their complex conjugates \( \overline{d}_j \). Suppose that \( \Pi_a < \Pi, \forall a \in [1, N] \) for some deterministic and positive function \( \Pi(z) \) that satisfies \( \frac{\sqrt{7} N}{\sqrt{N}} + \Psi^2 < \Psi \). Then,

\[
\left| \frac{\sqrt{7} N}{N^2} \sum_a d_a L_{aa} \right| < \Psi \Pi.
\]

As in Lemma B.4, it is suffices to show the following recursive moment estimate.

**Lemma C.2.** Fix a \( z \in D \). Suppose that the assumptions of Proposition B.4 hold. Then, for any fixed integer \( p \geq 1 \), we have

\[
\mathbb{E}[n^{(p,p)}] = \mathbb{E}[O(\Pi^2)n^{(p-1,p)}] + \mathbb{E}[\Psi^2\Pi^2]n^{(p-1,p-1)},
\]

where we defined

\[
n^{(k,i)} := \left( \frac{\sqrt{7} N}{N^2} \sum_a d_a L_{aa} \right)^k \left( \frac{\sqrt{7} N}{N^2} \sum_a d_a L_{aa} \right)^i.
\]

**Proof.** From the definition of \( n^{(p,p)} \) and the Stein’s lemma, we have that

\[
\mathbb{E}[n^{(p,p)}] = \frac{\sqrt{7} N}{N^2} \mathbb{E}\left[ \sum_a d_a W_{aa} G_{aa} n^{(p-1,p)} \right] + \frac{1}{N^2} \sum_a \mathbb{E}[\text{tr}(G)G_{aa} n^{(p-1,p)}] \\
= \frac{\sqrt{7} N}{N^2} \mathbb{E}\left[ \sum_a d_a W_{aa} G_{aa} n^{(p-1,p)} \right] + \frac{p-1}{N^2} \sum_a \mathbb{E}[d_a G_{aa} \frac{\partial n^{(1,0)}}{\partial W_{aa}} n^{(p-2,p)}] + \frac{p}{N^2} \sum_a \mathbb{E}[d_a G_{aa} \frac{\partial n}{\partial W_{aa}} n^{(p-1,p-1)}].
\]

The first term can be handed by the assumption in Proposition C.1 Remaining terms can be dealt with Lemma C.3 below.

**Lemma C.3.** Fix a \( z \in D \). Suppose that the assumptions of Proposition C.2 hold and let \( Q \) be an \((N \times N)\) matrix. Then we have

\[
\frac{\sqrt{7} N}{N^2} \sum_i \sum_k d_i G_{ki} \text{tr} \left( Q \frac{\partial W G}{\partial W_{ki}} \right) = O_\infty(\|Q\| \Psi^2 \Pi^2),
\]

\[
\frac{\sqrt{7} N}{N^2} \sum_i \sum_k d_i G_{ki} \text{tr} \left( Q \frac{\partial G}{\partial W_{ki}} \right) = O_\infty(\|Q\| \Psi^2 \Pi^2),
\]

and the same estimates hold if we replace the \( \partial W_{da} \) by \( \partial W_{ad} \).
Proof. Computing the derivative gives that
\[
\frac{1}{N^2} \sum_{i} \sum_{k} d_i G_{ki} \text{tr} \left( Q \frac{\partial (WG)}{\partial W_{ki}} \right) = \frac{1}{N^2} \sum_{i} \sum_{k} d_i G_{ki} \text{tr} \left( Q \frac{\partial W}{\partial W_{ki}} G \right) + \frac{1}{N^2} \sum_{i} \sum_{k} d_i G_{ki} \text{tr} \left( Q W \frac{\partial G}{\partial W_{ki}} \right).
\]
For the first term, we have that
\[
\frac{1}{N^2} \left| \sum_{i} \sum_{k} d_i G_{ki} \text{tr} \left( Q \frac{\partial W}{\partial W_{ki}} G \right) \right| = \frac{1}{N^2} \left| \sum_{i} \sum_{k} d_i G_{ki} (GQ)_{ik} \right| \leq \frac{1}{N^2} \sum_{i} \| Q \|_{\text{im}} \leq \frac{\| Q \|}{N^2} \sqrt{\Pi^2},
\]
where we used the Cauchy-Schwarz inequality and (C.1). The other term can be estimated as
\[
\frac{1}{N^2} \left| \sum_{i} \sum_{k} d_i G_{ki} \text{tr} \left( Q W \frac{\partial G}{\partial W_{ki}} \right) \right| \leq \frac{1}{N^2} \sum_{i} d_i e_i^* G(W G e_i G) \leq C \left( \frac{1}{N^2} \Pi^2 + \frac{1}{N} \right)
\]
where we used (B.23) and the Cauchy-Schwarz inequality. The last equation in (C.4) can be obtained in a similar way.

Proposition C.4. Fix a \( z \in \mathcal{D} \). Suppose the assumptions of Proposition B.8 hold. Let \( d_1, \ldots, d_N \in \mathbb{C} \) be possibly \( H \)-dependent quantities satisfying \( \max |d_i| < 1 \). Assume that for all \( i, j \in [1, N] \),
\[
\frac{t}{N} \sum_{k} \frac{\partial d_j}{\partial W_{ki}} e_i W G e_i = O(\Psi^2 \Pi^2),
\]
and the same bounds hold when the \( d_j \)'s are replaced by their complex conjugates \( \bar{d}_j \). Suppose that \( \Pi_a < \tilde{\Pi}, \forall a \in [1, N] \) for some deterministic and positive function \( \tilde{\Pi}(z) \) that satisfies \( \frac{i}{\sqrt{N^2 \Pi^2} + \Psi^2} < \tilde{\Pi} < \Psi \). Then,
\[
\left| \frac{t}{N} \sum_{a} d_a J_a \right| < \Psi \tilde{\Pi}.
\]

The proof is omitted since it is the same as that of Proposition C.3 except that we use Lemma C.6 below as an input, instead of Lemma C.5.

Lemma C.5. Fix a \( z \in \mathcal{D} \). Suppose that the assumptions of Proposition C.4 hold. Let \( Q \) be a matrix. Then we have
\[
\frac{t}{N^2} \sum_{i} \sum_{k} d_i (WG)_{ik} \text{tr} \left( Q \frac{\partial (WG)}{\partial W_{ik}} \right) = O(\|Q\| \Psi^2 \Pi^2),
\]
\[
\frac{t}{N^2} \sum_{i} \sum_{k} d_i (WG)_{ik} \text{tr}(QWG) \text{tr} \left( \frac{\partial G}{\partial W_{ik}} \right) = O(\|Q\| \Psi^2 \Pi^2),
\]
\[
\frac{t}{N^2} \sum_{i} \sum_{k} d_i (WG)_{ik} \text{tr}(G) \text{tr} \left( Q \frac{\partial (WG)}{\partial W_{ik}} \right) = O(\|Q\| \Psi^2 \Pi^2),
\]
and the same estimates hold if we replace the \( \partial W_{da} \) by \( \partial W_{ad} \).

Proof. Since the other terms can be shown in similar way, we only consider the first equation in (C.6). Computing the derivative, we have that
\[
\frac{t^2}{N^2} \sum_{i,k} d_i (WG)_{ik} \text{tr} \left( Q \frac{\partial (WG)}{\partial W_{ik}} \right) = \frac{t^2}{N^2} \sum_{i,k} d_a (WG)_{ak} (QGW)_{ki} - \frac{t^2}{N^2} \sum_{i,k} d_a (WG)_{ik} (GWQWG)_{ki} + \frac{t^2}{N^2} \sum_{a} d_a (WGQWG)_{aa} = O \left( \frac{\|Q\|}{N} \Pi^2 \right) + O \left( \frac{t^2}{N^2} \|Q\| (\Pi^2) \right) + O \left( \frac{t}{N^2} \Pi^2 (\Pi^2) \right)
\]
where we have used the Cauchy-Schwarz inequality, (B.23), and (C.4).
Now we turn to general averages of $Q_{aa}$, which is an analogue of \cite{S} Proposition 6.1:

**Proposition C.6.** Fix a $z \in \mathbb{D}$. Suppose the assumptions of Proposition B.8 hold. Let $d_1, \ldots, d_N \in \mathbb{C}$ be possibly $H$-dependent quantities satisfying $\max |d_i| \prec 1$. Assume that for all $i, j \in [1, N]$,

$$
\frac{1}{N} \sum_k \frac{\partial d_i}{\partial y_{ik}} e_i^k x_i G e_i = O(\Psi^2 \Pi^2),
\frac{1}{N} \sum_k \frac{\partial d_i}{\partial y_{ik}} e_i^k x_i \hat{g}_i = O(\Psi^2 \Pi^2),
$$

(C.10)

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

$$
\left| \frac{1}{N} \sum_a d_a Q_{aa} \right| \prec \Psi \Pi.
$$

Again, we first reduce the proof of Proposition C.6 to the corresponding recursive moment estimate.

**Lemma C.7.** Fix a $z \in \mathbb{D}$. Suppose that the assumptions of Proposition C.6 hold. Then, for any fixed integer $p \geq 1$, we have

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

(C.11)

where we defined

$$
m^{(k,l)} := \left( \frac{1}{N} \sum_a d_a Q_{aa} \right)^k \left( \frac{1}{N} \sum_a d_a Q_{aa} \right)^l.
$$

(C.12)

**Proof of Lemma C.7.** We first claim that if $|\Upsilon| \prec \hat{\Upsilon}(z)$ for a deterministic, positive function $\hat{\Upsilon} \leq \Psi(z)$, then

$$
\mathbb{E}[m^{(p,p)}] = \mathbb{E}[O(\hat{\Upsilon}^2 + O(\Psi \hat{\Upsilon}))m^{(p-1, p)}] + \mathbb{E}[O(\Psi^2 \hat{\Upsilon}^2)m^{(p-2, p)}] + \mathbb{E}[O(\Psi^2 \hat{\Upsilon}^2)m^{(p-1, p-1)}],
$$

(C.13)

with conventions $m^{(0,0)} = 1 = m^{(-1,1)}$. The estimate (C.13) can be proved using the methods from \cite{S} Lemma 6.2. The major difference from \cite{S} Lemma 6.2 arises when we differentiate entries or traces of $BG$ with respect to $g_i$'s, in which case additional terms involving derivatives of $WG$ appear due to (B.20). Since $U$ and $W$ are independent, these terms can be handled with estimates in Lemma B.7. Due to the similarity with Lemma 6.2 in \cite{S} we omit further details.

Next, we prove Lemma C.7 from (C.13). Firstly, Young’s and Markov’s inequalities provide that

$$
\left| \frac{1}{N} \sum_a d_a Q_{aa} \right| \prec \hat{\Upsilon}^2 + \Psi \hat{\Upsilon} + \Psi \hat{\Upsilon} \prec \Psi \hat{\Upsilon} + \Psi \hat{\Pi}.
$$

(C.14)

Now we note that

$$
\Upsilon = \frac{1}{N} \sum_a a_a Q_{aa} - \sqrt{t} \text{tr}(G) \text{tr}(W) + \frac{\sqrt{t}}{N} \sum_a \text{tr}((I - z + B)G) L_{aa} + \frac{t}{N} \sum_a \text{tr}(G) J_a.
$$

We can check that the weights for $L_{aa}$ and $J_a$ satisfy the assumptions in Propositions C.1 and C.4 respectively, hence we can apply them to obtain

$$
|\Upsilon| \prec \Psi \hat{\Upsilon} + \Psi \hat{\Pi} \prec 1 + \Psi \hat{\Pi} \prec N^{-1} \Psi \hat{\Pi} + \Psi \hat{\Pi}.
$$

(C.15)

Updating $\hat{\Upsilon}$ as the right hand side of (C.15) and iterating (C.13) repeatedly give $|\Upsilon| \prec \Psi \Pi$. Hence we finally choose $\hat{\Upsilon} = \Psi \Pi$ and use the assumption $\frac{1}{\sqrt{N}} + \Psi^2 \prec \Psi$ to conclude Lemma C.7. \qed
C.2 Optimal fluctuation averaging

In this section, we prove the following improved estimate for a specific linear combination of $Q_{aa}$'s. This result is an analogue of Proposition 7.1 and is used later to validate the assumption of Lemma 4.2 which leads to an estimate for $\Lambda$.

**Proposition C.8.** Fix a $z = E + i\eta \in \mathcal{D}_x$. Suppose that the assumptions of Proposition 4.8 hold. Suppose that $\Lambda < \tilde{\Lambda}$, for some deterministic and positive function $\tilde{\Lambda} \prec N^{-\epsilon/4}$, then

$$|SA_{i} + TA_{i}^2 + O(\Lambda^3)| \prec \frac{\sqrt{(Im \tilde{m} + \tilde{\Lambda})(|S| + \tilde{\Lambda})}}{N\eta} + \frac{1}{(N\eta)^2}, \quad i = A, B. \tag{C.16}$$

We first express the left-hand side of (C.16) in terms of weighted averages of $Q_{aa}$ and $L_{aa}$'s. To obtain such linear combination, we recall that the subordinate system $\Phi_{AB}$ defined in (3.3) vanishes at the point $(\omega_A(z), \omega_B(z), z)$.

As our final goal is to bound the difference between the approximate and genuine subordination functions, we evaluate the system at the point $(\omega_A^e(z), \omega_B^e(z), z)$;

$$\Phi_{A} := \Phi_{A}(\omega_A^e, \omega_B^e, z) \quad \text{and} \quad \Phi_{B} := \Phi_{B}(\omega_A^e, \omega_B^e, z).$$

From (B.26) and (B.7), we have that

$$\Lambda = 1/N \sum_{a} (d_{A,a} Q_{aa} + \sqrt{\frac{1}{c}} \tilde{\Phi}_{a}^{W} L_{aa}),$$

and

$$F_{A}(\omega_{A}) + \frac{1}{tr \ G} - t tr \ G = \left( -\frac{1}{m_{A}(\omega_{A})} \right) G,$$

where we defined

$$d_{A,a} := \left( -\frac{1}{m_{A}(\omega_{A})} \right) \frac{1}{(a_{a} - \omega_{A}^e) tr \ G}, \quad \tilde{\Phi}_{a}^{W} := \frac{1}{tr \ G},$$

and $d_{B,a}, \tilde{\Phi}_{a}^{W}$ symmetrically. On the other hand, we expand $\Phi_{A}$ and $\Phi_{B}$ around $(\omega_A, \omega_B, z)$ to obtain

$$\Phi_{A} = -\Lambda_{A} + (F_{A}^{e}(\omega_{A}) - 1) \Lambda_{A} + \frac{1}{2} F_{A}^{e}(\omega_{A}) \Lambda_{A}^2 + O(\Lambda_{A}^3),$$

and

$$\Phi_{B} = -\Lambda_{B} + (F_{B}^{e}(\omega_{B}) - 1) \Lambda_{B} + \frac{1}{2} F_{B}^{e}(\omega_{B}) \Lambda_{B}^2 + O(\Lambda_{B}^3).$$

Combining (C.17) and (C.18), we write

$$\Phi_{A} = \frac{1}{N} \sum_{a} (d_{A,a} Q_{aa} + \sqrt{\frac{1}{c}} \tilde{\Phi}_{a}^{W} L_{aa}) \quad \text{and} \quad \Phi_{B} = \frac{1}{N} \sum_{a} (d_{B,a} Q_{aa} + \sqrt{\frac{1}{c}} \tilde{\Phi}_{a}^{W} L_{aa}), \tag{C.19}$$

and (C.19) and (C.20), we get

$$Z_{A} := (F_{B}^{e}(\omega_{B}) - 1) \Phi_{A} + \Phi_{B} = SA_{i} + TA_{i}^2 + O((\Phi_{A}^e)^2) + O(\Phi_{A}^{e} L_{A}) + O(\Lambda_{A}^3). \tag{C.21}$$

Combining C.8 and C.21, we often need to apply Proposition G.6 with $\hat{\Pi}$ chosen to be the square root of the right-hand side of (C.16), i.e.,

$$\hat{\Pi}^2 = \frac{\sqrt{(Im \tilde{m} + \tilde{\Lambda})(|S| + \tilde{\Lambda})}}{N\eta} + \frac{1}{(N\eta)^2}. \tag{C.22}$$

Thus we need to prove that $\hat{\Pi}$ satisfies the assumptions in Proposition G.6. To this end, we claim that $|m_{H_{i}} - \tilde{m}| \prec \tilde{\Lambda} + \frac{\psi^2}{|m_{H_{i}}(z)|}$ when $\Lambda \prec \tilde{\Lambda} \prec N^{-\epsilon/4}$. First, observe from the definition of approximate subordination functions (3.4) that

$$|F_{H_{i}}(z) - F_{\tilde{H}_{i}}(z)| = |\omega_{A}^e(z) - \omega_{A}(z) + \omega_{B}^e(z) - \omega_{B}(z) + t \frac{1}{N m_{H_{i}}(z)} \sum_{a} L_{aa}| \prec \Lambda + \frac{\psi^2}{|m_{H_{i}}(z)|}. \tag{C.23}$$
where we applied Proposition C.1. On the other hand, recall from the definition of $F_t$ that

$$|F_h(z) - F_{h_0}(z)| = |m_t - \hat{m}| \left| \frac{1}{m_t \hat{m}} - t \right|. \tag{C.24}$$

Due to Lemmas A.3 and A.11, there exists $C > 0$ such that $|\hat{m}(z)| < C$ for any $z \in \mathcal{D}_r$. On the other hand, Cauchy-Schwarz inequality and compactness of $\mathcal{D}_r$ imply that there exists $c > 0$ satisfying

$$\int_{\mathbb{R}} \left| \frac{1}{|x-z|^2} d\mu(x) - \int_{\mathbb{R}} \frac{1}{|x-z|^2} d\hat{\mu}(x) \right|^2 > c \tag{C.25}$$

uniformly on $\mathcal{D}_r$, so that for another constant $c' > 0$ we get $c' < 1 - t|\hat{m}(z)|^2$. Thus we have

$$c'|m_H - \hat{m}| - t|m_t - \hat{m}|^2 \leq |\hat{m} m_{H_t}| |F_{H_t} - F_{\hat{m}}| < \Lambda + \Psi^2 + |m_{H_t} - \hat{m}| \Lambda. \tag{C.26}$$

Due to the assumption $\Lambda_s < N^{-\sigma/4}$, we can solve this quadratic inequality to obtain $|m_{H_t} - \hat{m}| < \Lambda + \Psi^2$. Along with Proposition A.11, we can see the validity of $\hat{\Pi}$ for Proposition C.9.

Next, we prove Proposition C.8 assuming the validity of the following estimate for $Z_i$.

**Proposition C.9.** Fix $z \in \mathcal{D}_r$. Suppose that the assumptions of Proposition B.8 hold and that $\Lambda(z) < \Lambda(z)$ for some deterministic and positive function $\Lambda(z) \leq N^{-\sigma/4}$. Choose $\Pi(z)$ as in (C.22). Then,

$$|Z_A| < \hat{\Pi}^2, \quad |Z_B| < \hat{\Pi}^2. \tag{C.27}$$

**Proof of Proposition C.9.** By definition of $\omega^i$, the chain rule and (B.35), it is easy to check that $\partial_{\omega^i}$ and $\partial_{\omega^j}$, satisfy the assumptions in Proposition C.6 and C.1 respectively for $\omega^A$ and $\omega^i$, $i = 1, 2, \ldots, N$. Hence we have that

$$\Phi^i | < \Psi \hat{\Pi} \quad i = A, B. \tag{C.28}$$

Combining this estimate, equation (C.21) and Proposition C.8 gives that

$$|\mathcal{S} \Lambda + \mathcal{C} \Lambda^2 + O(\Lambda^3) | < \hat{\Pi}^2 + \Psi \hat{\Pi} \Lambda \quad i = A, B. \tag{C.29}$$

The definition of $\hat{\Pi}$ implies that the second term on the right hand side can be absorbed into the first term. Thus we have Proposition C.8.

As in the previous proofs, the proof of Proposition C.8 reduces to the following recursive moment.

**Lemma C.10.** Fix $z \in \mathcal{D}_r$. Suppose that the assumptions of Proposition C.8 hold. For any fixed $p \geq 1$, we have

$$E[L(z)] = E[O(\hat{\Pi}^2) L^{(p-1,p)}] + E[O(\hat{\Pi}) L^{(p-2,p)}] + E[O(\hat{\Pi}^4) L^{(p-1,p-1)}], \tag{C.30}$$

where we denote

$$\Lambda^{(k,l)} = \Lambda^{(k,l)}(z) := Z_{\Lambda} Z_{\Lambda}, \quad k, l \in \mathbb{N}. \tag{C.31}$$

with conventions $L^{(0,0)} = 1$ and $L^{(-1,1)} = 0$.

**Proof of Lemma C.10.** Recall that

$$E[L^{(p,p)}] = E[Z_A L^{(p-1,p)}] = \frac{1}{N} \sum_a E[\partial_{a_A} Q_{aa} L^{(p-1,p)}] + \frac{1}{N} \sum_a E[\partial_{a_B} Q_{aa} L^{(p-1,p)}]$$

$$+ \sqrt{\frac{1}{N}} \sum_a E[\partial_{W_{a_a}} L_{aa} L^{(p-1,p)}] \tag{C.32}$$

We follow the strategy of proof of Lemma 7.3 in [8] so that the main task of the proof is estimating the terms including the derivatives of $Z_A$ or $Z_{\Lambda}$ (cf. Lemma 7.4 in [8]). Due to similarity, we only mention that the extra gain for the estimate comes from the fact that

$$\frac{\partial Z_A}{\partial g_{ab}} = (S_{AB} + O(\Lambda)) \frac{\partial \omega_A}{\partial g_{ab}} + O(\Lambda) \frac{\partial \omega_B}{\partial g_{ab}}, \tag{C.33}$$

which directly follows from the definition of $\Phi_{AB}$. Replacing derivatives with respect to $g_{ab}$ by $W_{ab}$ in (C.33) proves the estimate for the third and fourth terms. \qed
D Proofs of Theorem B.1 and Lemma 4.5

In this section we prove Theorem B.1 and Lemma 4.5. The proof of Theorem B.1 consists of two parts, weak local law and strong local law. The former implies that the assumptions in Proposition B.3 hold uniformly true on $D$ and the latter proves Theorem B.1. The proof of Lemma 4.5 is presented at the end of this section.

D.1 Weak local law

In this section we establish a weaker version of local law that holds uniformly on $D$. As in [8], this weak law will serve as an input for the proof of strong local law, Theorem B.1. We state the weak law in the following theorem.

**Theorem D.1.** Suppose that Assumption 2.3 holds. Then, for all $i, j \in [1, N]$, we have

$$|P_{ij}(z)| \prec \Pi_i(z) + \Pi_j(z),$$

$$|K_{ij}(z)| \prec \Pi_i(z) + \Pi_j(z),$$

$$\Lambda_L(z) \prec \Psi(z),$$

$$\Lambda_T \prec \Psi(z),$$

uniformly in $z \in D$. In addition, we have

$$\Lambda_e \prec \frac{1}{N^\eta \eta_s^{1/3}},$$

$$\Lambda \prec \frac{1}{N^\eta \eta_s^{1/3}},$$

(D.1)

uniformly in $z \in D$. The same statements hold for analogous quantities with roles of $(A, B)$, $(U, U^*)$, and $(W, W)$ interchanged.

Before we prove the theorem, we introduce the following lemma which enables us to convert the bound in Proposition C.8 to bounds for $\Lambda$. We omit its proof for it is exactly the same as [8] Lemma 8.2.

**Lemma D.2.** Fix $z \in D$. Let $\epsilon \in (0, \frac{\sigma}{100})$ and $k \in (0, 1]$. Let $\tilde{\Lambda} \equiv \tilde{\Lambda}(z)$ be some deterministic control parameter satisfying $\tilde{\Lambda} \leq N^{-\sigma / 4}$. Suppose that $\Lambda(z) \leq \tilde{\Lambda}(z)$ and

$$|\Sigma A_C + T_C \Lambda_C^2 + O(\Lambda_C^2)| \leq N^2 \frac{|\Sigma| + \tilde{\Lambda}}{(N \eta)^k}, \quad C = A, B$$

hold on some event $\Omega(z)$. Then there exists a constant $C > 0$ such that for sufficiently large $N$, the following hold:

(i) If $\sqrt{\kappa + \eta} > N^{-\tilde{\Lambda}}$, there is a sufficiently large constant $K_0 > 0$ independent of $z$, such that

$$\mathbb{1} \left( \Lambda \leq \frac{|\Sigma|}{K_0} \right) \left| \Lambda_A \right| \leq \left( N^{-2\tilde{\Lambda}} \Lambda + \frac{N^2}{(N \eta)^k} \right), \quad \mathbb{1} \left( \Lambda \leq \frac{|\Sigma|}{K_0} \right) \left| \Lambda_B \right| \leq \left( N^{-2\tilde{\Lambda}} \Lambda + \frac{N^2}{(N \eta)^k} \right)$$

on $\Omega(z)$,

(D.2)

where $\mathbb{1}$ denotes the indicator function.

(ii) If $\sqrt{\kappa + \eta} \leq N^{-\tilde{\Lambda}}$, we have

$$\left| \Lambda_A \right| \leq \left( N^{-\tilde{\Lambda}} \Lambda + \frac{N^2}{(N \eta)^k} \right), \quad \left| \Lambda_B \right| \leq \left( N^{-\tilde{\Lambda}} \Lambda + \frac{N^2}{(N \eta)^k} \right)$$

on $\Omega(z)$.

For $z \in D$ and $\delta, \delta' \in [0, 1]$, we define the event

$$\Theta(z, \delta, \delta') := \left\{ \Lambda_d \leq \delta, \tilde{\Lambda}_d(z) \leq \delta, \Lambda(z) \leq \delta, \Lambda_d(z) \leq \delta', \Lambda_T(z) \leq \delta', \tilde{\Lambda}_T(z) \leq \delta' \right\}.$$

In addition, we decompose the domain $D$ into the following disjoint parts:

$$D_\gamma := \left\{ z \in D : \sqrt{\kappa + \eta} > \frac{N^2}{(N \eta)^k} \right\}, \quad D_\delta := \left\{ z \in D : \sqrt{\kappa + \eta} \leq \frac{N^2}{(N \eta)^k} \right\}.$$

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

$$\Theta_\gamma(z, \delta, \delta', \epsilon') := \left\{ \Lambda_d(z) \leq \delta, \tilde{\Lambda}_d(z) \leq \delta, \Lambda(z) \leq \min\{\delta, N^{-\epsilon'}|\Sigma|\}, \Lambda \leq \delta', \Lambda_T(z) \leq \delta', \tilde{\Lambda}_T(z) \leq \delta' \right\}.$$

With these notations, the following lemma lets us use a bootstrapping argument;
Lemma D.3. For any fixed $z \in \mathcal{D}$, any $\epsilon \in (0, \frac{1}{10}]$ and any $D > 0$, there exists a positive integer $N(D, \epsilon)$ and an event $\Omega(z) := \Omega(z, D, \epsilon)$ with
\[
\mathbb{P}(\Omega(z)) \geq 1 - N^{-D}, \quad \forall N \geq N(D, \epsilon),
\]
such that the following hold:

(i) If $z \in \mathcal{D}_>$, we have
\[
\Theta_\epsilon \left( z, \frac{N^{3\epsilon}}{(N \eta)^{1/3}}, \frac{N^{3\epsilon}}{\sqrt{N \eta}}, \frac{\epsilon}{10} \right) \cap \Omega(z) \subset \Theta_\epsilon \left( z, \frac{N^{5\epsilon}}{(N \eta)^{1/3}}, \frac{N^{5\epsilon}}{\sqrt{N \eta}}, \frac{\epsilon}{2} \right).
\]

(ii) If $z \in \mathcal{D}_<$, we have
\[
\Theta \left( z, \frac{N^{3\epsilon}}{(N \eta)^{1/3}}, \frac{N^{3\epsilon}}{\sqrt{N \eta}} \right) \cap \Omega(z) \subset \Theta \left( z, \frac{N^{5\epsilon}}{(N \eta)^{1/3}}, \frac{N^{5\epsilon}}{\sqrt{N \eta}} \right).
\]

Proof. As in [8, Lemma 8.3], the proof of Lemma D.3 requires quantitative versions of each of moment estimates in previous sections. To be specific, we need their counterparts which do not depend on the probabilistic input [B.32]. In this sense the estimates for entrywise local laws, Lemmas B.2, B.4, and estimates in previous sections. To be specific, we need their counterparts which do not depend on the

\[
\text{Lemma D.3.}
\]

For any fixed $D, \epsilon > 0$, there exists a positive integer $N(D, \epsilon)$ and an event $\Omega(z) := \Omega(z, D, \epsilon)$ with
\[
\mathbb{P}(\Omega(z)) \geq 1 - N^{-D}, \quad \forall N \geq N(D, \epsilon),
\]
such that the following hold:

(i) If $z \in \mathcal{D}_>$, we have
\[
\Theta_\epsilon \left( z, \frac{N^{3\epsilon}}{(N \eta)^{1/3}}, \frac{N^{3\epsilon}}{\sqrt{N \eta}}, \frac{\epsilon}{10} \right) \cap \Omega(z) \subset \Theta_\epsilon \left( z, \frac{N^{5\epsilon}}{(N \eta)^{1/3}}, \frac{N^{5\epsilon}}{\sqrt{N \eta}}, \frac{\epsilon}{2} \right).
\]

(ii) If $z \in \mathcal{D}_<$, we have
\[
\Theta \left( z, \frac{N^{3\epsilon}}{(N \eta)^{1/3}}, \frac{N^{3\epsilon}}{\sqrt{N \eta}} \right) \cap \Omega(z) \subset \Theta \left( z, \frac{N^{5\epsilon}}{(N \eta)^{1/3}}, \frac{N^{5\epsilon}}{\sqrt{N \eta}} \right).
\]

we first derive consequences of (D.3). Since
\[
\varphi(\Gamma_{ij}^a) = 1 = \varphi(\Gamma_{ij}) \quad \text{on } \Theta \left( z, \frac{N^{3\epsilon}}{(N \eta)^{1/3}}, \frac{N^{3\epsilon}}{\sqrt{N \eta}} \right),
\]
we may apply Markov’s inequality to (D.3) to find an event $\Omega_1 \equiv \Omega_1(z, D, \epsilon)$ so that
\[
|J_i| \leq N^{\sigma/4} \Psi, \quad |L_{ij}| \leq N^{\sigma/4} \Psi, \quad |P_{ij}| \leq N^{\sigma/4} \Psi, \quad |K_{ij}| \leq N^{\sigma/4} \Psi
\]
are all true on the event $\Theta \left( z, \frac{N^{3\epsilon}}{(N \eta)^{1/3}}, \frac{N^{3\epsilon}}{\sqrt{N \eta}} \right) \cap \Omega_1$, and that
\[
\mathbb{P}(\Omega_1) \leq \frac{1}{10} N^{-D}.
\]

Then we follow the proof of Proposition B.8 to obtain that
\[
\Lambda_\mu(z) \leq \frac{N^{\sigma/2}}{\Psi}, \quad \Lambda_\eta(z) \leq N^{\sigma/2} \Psi, \quad \Lambda_T(z) \leq N^{\sigma/2} \Psi, \quad \Lambda_T(z) \leq N^{\sigma/2} \Psi, \quad |\gamma(z)| \leq N^{\sigma/2} \Psi
\]
hold on $\Theta \left( z, \frac{N^{3\epsilon}}{(N \eta)^{1/3}}, \frac{N^{3\epsilon}}{\sqrt{N \eta}} \right) \cap \Omega_1(z)$.

Now we present the quantitative versions of estimates in Section C.2 with weaker bounds. Specifically, we consider the following two quantities;
\[
\tilde{\eta}^{(\rho, \sigma)} := \left( \frac{1}{N} \sum_i \sigma_{A,i} L_{ii} \varphi(\Gamma_{i0}) \right)^{\rho} \left( \frac{1}{N} \sum_i \sigma_{A,i} L_{ii} \varphi(\Gamma_{i0}) \right)^{\sigma},
\]
\[
\tilde{\eta}^{(\rho, \sigma)} := \left( \frac{1}{N} \sum_i \sigma_{A,i} Q_{ii} \varphi(\Gamma_{ii}) \varphi(\Gamma) \right)^{\rho} \left( \frac{1}{N} \sum_i \sigma_{A,i} Q_{ii} \varphi(\Gamma_{ii}) \varphi(\Gamma) \right)^{\sigma},
\]
\]
where $\gamma_{ij}^a$ and $\Gamma_{ij}$ are as in [B.31, B.32] and $\Gamma$ is defined as
\[
\Gamma := (c \text{Im} \tilde{m} + \tilde{\Lambda})^{-2} (|A_t|^2 + |A_t|^2) + \left( \frac{N^{5\epsilon}}{(N \eta)^{1/3}} \right)^{-2} |\gamma|^2 + \left( \frac{N^{5\epsilon}}{\sqrt{N \eta}} \right)^{-1} \frac{1}{N} \sum_a (|T_{ij}|^2 + N^{-1})^{1/2},
\]
49
for some sufficiently small constant $c > 0$. In the rest of the proof, we choose

$$\tilde{\Lambda}(z) = \frac{N^{3\epsilon}}{(N\eta)^{1/3}},$$

and define $\hat{\Pi}$ as in (C.22).

Following the same calculations as in Proposition (C.4) but taking Lemma (B.5) as an additional input, we find that

$$E[\hat{i}^{(p,p)}] \leq N^{\epsilon_1} \psi^{4p}.$$

In addition, due to our choice of $\Lambda$, we have

$$\psi^2 = \frac{1}{N\eta} \leq \frac{1}{(N\eta)^{-2/3}} \leq \hat{\Pi}.$$

Now for the average of $Q_{ij}$’s, we follow the exact same argument as in Section C of [8] to prove that

$$E[\hat{m}^{(p,p)}] \leq N^{\epsilon_1} \hat{\Pi}^{2p}.$$

Repeating the same arguments with averages of $L_{aa}$ and $Q_{aa}$ and applying Markov’s inequality, since $\varphi(\Gamma) = \varphi(\Gamma_{ii}) = \varphi(\Gamma_{ia}) = 1$ on the event $\Theta \left( z N^{3\epsilon} \frac{N^{3\epsilon}}{(N\eta)^{1/3}}, \frac{N^{3\epsilon}}{\sqrt{N\eta}} \right)$, there exists an event $\Omega_2 \equiv \Omega_2(z, D, \epsilon)$ such that

$$\Phi^c \leq N^{\epsilon/2} \hat{\Pi}, \quad \text{on } \Theta \left( z N^{3\epsilon} \frac{N^{3\epsilon}}{(N\eta)^{1/3}}, \frac{N^{3\epsilon}}{\sqrt{N\eta}} \right) \cap \Omega_2(z)$$

and $P[\Omega_2(z)] \geq 1 - \frac{N^{-1}}{10}$. On this intersection of events, we further have from (C.10) that

$$|SA_n + T_i A_n^* + O(|A_n^*|^2)| \leq C N^{\epsilon/3} \hat{\Pi} \leq N^{\epsilon} \frac{|S| + \hat{\Lambda}}{(N\eta)^{1/3}}, \quad \text{if } A, B, i = A, B,$$

where we used the definition of $\hat{\Lambda}$ in the last inequality.

Taking $\Omega \equiv \Omega(z, D, \epsilon) = \Omega_1 \cap \Omega_2$, we find that the assumptions of Lemma (D.2) are satisfied with $k = 1/3$. The rest of the proof of Lemma (D.3) are identical to that of Lemma 8.3 in [8].

With Lemma (D.3) we now prove Theorem (D.1) by using a continuity argument.

**Proof of Theorem (D.1).** To prove Theorem (D.1), we modify that of Theorem 8.1 in [8]. Specifically, we first prove the result when $\eta = \eta_M \sim 1$, and then use Lemma (D.3) and a lattice continuity argument to gradually decrease $\eta$ until we reach the optimal regime $\eta = N^{-1+\sigma}$. Since the bootstrapping part of the proof is identical to that of Theorem 8.1 in [8], we focus on the first part. That is, we prove that there exists a sufficiently large constant $\eta_M > 0$ so that the conclusions of Theorem (D.1) hold uniformly for $z = E + i\eta_M \in D$.

Following (8.35) in [8], we can prove that $Q_{ij}$ as a function of $U$ is Lipschitz continuous with respect to the Hilbert-Schmidt norm $\|X\|_2 = \sqrt{\text{Tr} XX^*}$ with the Lipschitz constant bounded by $C\eta^{-2}$. Since the constant $C$ here can be chosen independent of $W$, applying Gromov-Milman concentration inequality (see Corollary 4.4.28 of [2]) to $Q_{ij}$ yields

$$|Q_{ij}(E + i\eta) - E[Q_{ij}(E + i\eta)]|W| \leq \frac{1}{\sqrt{N\eta^2}}$$

whenever $\eta \geq \eta_M$. Furthermore, using the invariance of the Haar measure we can check the identity

$$E[\hat{BG} \otimes G - G \otimes \hat{BG}|W] = 0.$$

Taking the $(i, j)$-th entry for the first component and the normalized trace for the second component in the tensor product, we have

$$E[Q_{ij}] = E[(\hat{BG})_{ij} \text{ tr } G - G_{ij} \text{ tr } \hat{BG}|W] = 0. \quad (D.8)$$
Following the same proof as that of (8.38) in [3], we can extend the bound to the whole domain \( \eta \geq \eta_M \) (enlarging \( \eta_M \) if necessary):

\[
\sup_{z: \|z\| \geq \eta_M} |Q_{ij}(z)| \prec \frac{1}{\sqrt{N}}, \quad \forall i, j \in [1, N]. \tag{D.9}
\]

Similarly, applying Proposition 2.3.3 in [2] to \( L_{ab} \) as a function of \( W \), we get

\[
\sup_{z: \|z\| \geq \eta_M} |L_{ij}(z)| \prec \frac{1}{\sqrt{N}}, \quad \forall i, j \in [1, N].
\]

In addition, using that \( \|H_t\| \leq \|A\| + \|B\| + t\|W\| \leq C \) with high probability and that \( \text{tr} \, \tilde{B} = \text{tr} \, B = 0 \), we have, for \( z = E + i\eta \) with fixed \( E \) and any \( \eta \geq \eta_M \), the expansions

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

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

\[
\omega_A^\epsilon = z - \frac{\text{tr}(\tilde{B}G + \sqrt{W}G)}{\text{tr}(G)} = z + O\left(\frac{1}{\eta_M}\right), \quad z = E + i\eta_M. \tag{D.10}
\]

Using the identity \((\tilde{B}G)_{ij} = \delta_{ij} - (a_i - z)G - \sqrt{t}WG)_{ij}\), we can rewrite (D.9) as

\[
(\delta_{ij} - (a_i - z)G - \sqrt{t}WG)_{ij} \text{tr} \, G + \sqrt{t} \text{tr}(WG)G_{ij} - \sqrt{t}(WG)_{ij} \text{tr}(G) = O\left(\frac{1}{\sqrt{N}}\right), \quad z = E + i\eta_M.
\]

Thus we have

\[
\Lambda_A^\epsilon(z) \prec \frac{1}{\sqrt{N}}, \quad \Lambda_L(z) \prec \frac{1}{\sqrt{N}}, \quad z = E + i\eta_M. \tag{D.11}
\]

Taking the average of diagonal terms in (D.11) yields

\[
\sup_{z: \|z\| \geq \eta_M} |m_{H_t}(z) - m_A(\omega_A^\epsilon(z))| \prec \frac{1}{\sqrt{N}}, \quad \sup_{z: \|z\| \geq \eta_M} |m_{H_t}(z) - m_B(\omega_B^\epsilon(z))| \prec \frac{1}{\sqrt{N}}
\]

where in the large \( z \) regime these bounds even hold deterministically. This gives the system

\[
\sup_{z: \|z\| \geq \eta_M} |\Phi_A(\omega_A^\epsilon(z), \omega_B^\epsilon(z), z)| \prec \frac{1}{\sqrt{N}}, \quad \sup_{z: \|z\| \geq \eta_M} |\Phi_B(\omega_A^\epsilon(z), \omega_B^\epsilon(z), z)| \prec \frac{1}{\sqrt{N}}. \tag{D.12}
\]

We regard (D.12) as a perturbation of \( \Phi_{A_B}(\omega_A(z), \omega_B(z), z) = 0 \), whose stability in the macroscopic regime is provided in Lemma A.12. Since (D.12) and (D.10) hold for sufficiently large \( \eta_M \), Lemma A.12 implies that

\[
|\Lambda_\tau(z)| = |\omega_\tau^\epsilon(z) - \omega_\tau(z)| \prec \frac{1}{\sqrt{N}}, \quad \tau = A, B, \quad z = E + i\eta_M,
\]

taking larger \( \eta_M > 1 \) if necessary. Thus we have

\[
|\Lambda_\tau(E + i\eta_M)| \leq |\Lambda_\tau^\epsilon(E + i\eta_M)| + |\Lambda_\tau(E + i\eta_M)| \prec \frac{1}{\sqrt{N}}, \tag{D.13}
\]

for any fixed \( E \in \mathbb{R} \). Using the bound \( \|G\| \leq \frac{1}{\eta_M} \) and the inequality \( |x^*Gy| \leq \|G\||x||y| \), we also get

\[
\Lambda_T(E + i\eta_M) \leq \frac{1}{\eta_M},
\]

for any fixed \( E \in \mathbb{R} \). Hence we observe that the assumptions in Proposition B.8 are satisfied so that we have, for any fixed \( E \in \mathbb{R} \), that

\[
|\Lambda_T(E + i\eta_M)| \prec \frac{1}{\sqrt{N}}, \quad |\tilde{\Lambda}_T(E + i\eta_M)| \prec \frac{1}{\sqrt{N}}. \tag{D.14}
\]
Also, note that \( E + i \eta M \in \mathcal{D} \) and \(|S(E + i \eta)| \geq 1\) for any fixed \( E \). Hence \( \Lambda(E + i \eta M) \prec N^{-1} |S(E + i \eta M)| \). From (D.13), we have

\[
\Lambda(E + i \eta M) \prec \frac{1}{\sqrt{N}}.
\]

Combining (D.13), (D.11), (D.14) and (D.15) with the fact \( \Lambda \prec N^{-1} |S(E + i \eta M)| \), we see that

\[
\Theta \left( E + i \eta M, \frac{N^{3\epsilon}}{N^{1/2}}, \frac{N^{3\epsilon}}{\sqrt{N}}, \frac{\epsilon}{10} \right) \geq 1 - N^{-D},
\]

for all \( E \) and \( N \geq N_0(D, \epsilon) \) with some sufficiently large \( N_0(D, \epsilon) \). This concludes the proof of Theorem D.1.

### D.2 Strong local law

In this section, we provide the strong local law, Theorem B.1.

#### Proof of Theorem B.1

We first prove the bound

\[
\Lambda(z) \prec \frac{1}{N^\eta}.
\]

Now that we have the weak local law Theorem D.1, the probabilistic assumptions in B.8 hold uniformly on the domain \( \mathcal{D} \). Thus the conclusion of Proposition C.8 holds true uniformly on \( \mathcal{D} \). That is,

\[
\left| \sum A + T \Lambda^2 + O(\Lambda^3) \right| \prec \frac{(\text{Im} \hat{m} + \hat{\Lambda})(|S| + \hat{\Lambda})}{N^\eta} + \frac{1}{(N^\eta)^2}, \quad \text{for } \Lambda(z) \prec \hat{\Lambda}.
\]

holds uniformly in \( \mathcal{D} \). Furthermore, since \( \text{Im} \hat{m} \leq C|S| \), we find that the assumptions of Lemma D.3 hold true for the choice \( k = 1 \) as long as \( \hat{\Lambda} \geq (N^\eta)^{-1} \) and \( \Lambda(z) \prec \hat{\Lambda} \).

As in the proof of Theorem 2.5 in [8], we use the bootstrapping argument for \( \hat{\Lambda} \) applying Lemma D.3 with \( k = 1 \). The initial choice is \( \hat{\Lambda}(z) = N^3 \left( N^\eta \right)^{-1/3} \), which is guaranteed by Theorem D.1, and we use the same argument as in [8] to iteratively improve the bound until we have (D.17).

Next, we prove Theorem B.1. Firstly, the averaged local law (B.1) is a consequence of Proposition C.6 and (D.17). Secondly for the entrywise local law (B.2), note that (D.17) implies

\[
\Pi_i(z) \prec \frac{\sqrt{\text{Im}(GW_i^*)^T(z) + \Lambda_i(z)}}{N^\eta} + \frac{1}{N^\eta}.
\]

Similarly, we have

\[
\Pi_i^W(z) = \frac{\sqrt{\text{Im}(GW_i^*)^T(z)}}{N^\eta} \prec \frac{\sqrt{|J_i| + |L_i| + \text{Im} \hat{m}(z) + \Lambda_i(z)}}{N^\eta} + \frac{1}{N^\eta}.
\]

Then (B.2) follows from the proof of Proposition B.8 once we replace the rough estimate \( \Psi \) with (D.19) and (D.20).

We conclude this section presenting the proof of Lemma 4.5.

#### Proof of Lemma 4.5

Using Proposition C.8 and Theorem B.1 in place of [8] Proposition 7.1 and Theorem 2.5, the following variant of Lemma 4.5 can be proved using the exact same proof as [8] Theorem 2.6:

\[
\max_{1 \leq i \leq N} |\lambda_i - \hat{\gamma}_i| \prec i^{-\frac{2}{3}} N^{-\frac{2}{3}},
\]

where \( \hat{\gamma}_i \)'s are the \( N \)-quantiles of \( \hat{\mu} \). Thus it suffices to prove

\[
\max_{1 \leq i \leq N} |\hat{\gamma}_i - \gamma_i| \prec i^{-1/3} N^{-2/3},
\]

where proof is analogous to that of [8] Lemma 3.14. We omit the detail due to similarity.
References

[1] A. Ahn. Airy point process via supersymmetric lifts, 2020.

[2] G. W. Anderson, A. Guionnet, and O. Zeitouni. An introduction to random matrices, volume 118 of Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 2010.

[3] J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. Ann. Probab., 33(5):1643–1697, 2005.

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

[5] Z. Bao, L. Erdős, and K. Schnelli. Convergence rate for spectral distribution of addition of random matrices. Adv. Math., 319:251–291, 2017.

[6] Z. Bao, L. Erdős, and K. Schnelli. Local law of addition of random matrices on optimal scale. Comm. Math. Phys., 349(3):947–990, 2017.

[7] Z. Bao, L. Erdős, and K. Schnelli. On the support of the free additive convolution. J. Anal. Math., 142(1):323–348, 2020.

[8] Z. Bao, L. Erdős, and K. Schnelli. Spectral rigidity for addition of random matrices at the regular edge. Journal of Functional Analysis, 279(7):108639, 2020.

[9] Z. Bao, K. Schnelli, and Y. Xu. Central Limit Theorem for Mesoscopic Eigenvalue Statistics of the Free Sum of Matrices. International Mathematics Research Notices, 09 2020. rnaa210.

[10] S. T. Belinschi. A note on regularity for free convolutions. Ann. Inst. Henri Poincaré Probab. Stat., 42(5):635–648, 2006.

[11] S. T. Belinschi. The Lebesgue decomposition of the free additive convolution of two probability distributions. Probab. Theory Related Fields, 142(1-2):125–150, 2008.

[12] S. T. Belinschi. $L^\infty$-boundedness of density for free additive convolutions. Rev. Roumaine Math. Pures Appl., 59(2):173–184, 2014.

[13] S. T. Belinschi and H. Bercovici. A new approach to subordination results in free probability. J. Anal. Math., 101:357–365, 2007.

[14] S. T. Belinschi, H. Bercovici, M. Capitaine, and M. Février. Outliers in the spectrum of large deformed unitarily invariant models. Ann. Probab., 45(6A):3571–3625, 2017.

[15] P. Bourgade, K. Mody, and M. Pain. Optimal local law and central limit theorem for $\beta$-ensembles, 2021.

[16] Z. Che and B. Landon. Local spectral statistics of the addition of random matrices. Probab. Theory Related Fields, 175(1-2):579–654, 2019.

[17] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. Spectral statistics of Erdős-Rényi Graphs II: Eigenvalue spacing and the extreme eigenvalues. Comm. Math. Phys., 314(3):587–640, 2012.

[18] V. Kargin. A concentration inequality and a local law for the sum of two random matrices. Probab. Theory Related Fields, 154(3-4):677–702, 2012.

[19] V. Kargin. An inequality for the distance between densities of free convolutions. Ann. Probab., 41(5):3241–3260, 2013.

[20] B. Landon and H.-T. Yau. Edge statistics of dyson brownian motion. 2017.

[21] J. O. Lee and K. Schnelli. Local deformed semicircle law and complete delocalization for Wigner matrices with random potential. J. Math. Phys., 54(10):103504, 62, 2013.

[22] J. O. Lee and K. Schnelli. Edge universality for deformed Wigner matrices. Rev. Math. Phys., 27(8):1550018, 94, 2015.

[23] J. O. Lee and K. Schnelli. Tracy-Widom distribution for the largest eigenvalue of real sample covariance matrices with general population. Ann. Appl. Probab., 26(6):3786–3839, 2016.

[24] J. O. Lee and K. Schnelli. Local law and Tracy-Widom limit for sparse random matrices. Probab. Theory Related Fields, 171(1-2):543–616, 2018.
[25] J. O. Lee, K. Schnelli, B. Stetler, and H.-T. Yau. Bulk universality for deformed Wigner matrices. *Ann. Probab.*, 44(3):2349–2425, 2016.

[26] J. O. Lee and J. Yin. A necessary and sufficient condition for edge universality of Wigner matrices. *Duke Math. J.*, 163(1):117–173, 2014.

[27] E. S. Meckes and M. W. Meckes. Spectral measures of powers of random matrices. *Electron. Commun. Probab.*, 18:no. 78, 13, 2013.

[28] L. Pastur and V. Vasilchuk. On the law of addition of random matrices: covariance and the central limit theorem for traces of resolvent. In *Probability and mathematical physics*, volume 42 of *CRM Proc. Lecture Notes*, pages 399–416. Amer. Math. Soc., Providence, RI, 2007.

[29] A. Soshnikov. Universality at the edge of the spectrum in Wigner random matrices. *Comm. Math. Phys.*, 207(3):697–733, 1999.

[30] T. Tao and V. Vu. Random matrices: universality of local eigenvalue statistics up to the edge. *Comm. Math. Phys.*, 298(2):549–572, 2010.

[31] C. A. Tracy and H. Widom. Level-spacing distributions and the Airy kernel. *Comm. Math. Phys.*, 159(1):151–174, 1994.

[32] C. A. Tracy and H. Widom. On orthogonal and symplectic matrix ensembles. *Comm. Math. Phys.*, 177(3):727–754, 1996.

[33] D. Voiculescu. Addition of certain noncommuting random variables. *J. Funct. Anal.*, 66(3):323–346, 1986.

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