Local laws for multiplication of random matrices and spiked invariant model

Xiucai Ding $^{*1}$ and Hong Chang Ji $^{12}$

$^1$Department of Statistics, University of California, Davis
$^2$Department of Mathematical Sciences, KAIST

Abstract

High dimensional Haar random matrices are common objects in modern statistical learning theory. We consider the random matrix model $A^{1/2}UBU^*A^{1/2}$, where $A$ and $B$ are two $N \times N$ positive definite matrices satisfying some regularity conditions and $U$ is either an $N \times N$ Haar unitary or orthogonal random matrix. On the macroscopic scale, it is well-known that the empirical spectral distribution (ESD) of the above model is given by the free multiplicative convolution of the ESDs of $A$ and $B$, denoted as $\mu_A \boxtimes \mu_B$, where $\mu_A$ and $\mu_B$ are the ESDs of $A$ and $B$, respectively.

In this paper, motivated by applications in statistical learning theory, we systematically study the above random matrix model. We first establish the local laws both near and far away from the rightmost edge of the support of $\mu_A \boxtimes \mu_B$. We also prove eigenvalue rigidity and eigenvector delocalization. Then we study the spiked invariant model by adding a few eigenvalues that are detached from the bulks of the spectrums to $A$ or $B$. We derive the convergent limits and rates for the outlier eigenvalues and their associated generalized components under general and weak assumptions. We believe all these results are optimal up to an $N^\epsilon$ factor, for some small $\epsilon > 0$.

1 Introduction

Large dimensional random matrices play important roles in high dimensional statistics and machine learning theory. Given a data matrix $Y$, researchers are interested in understanding the behavior of the singular values and vectors of $Y$, or equivalently, the eigenvalues and eigenvectors of $YY^*$ and $Y^*Y$. In particular, it has drawn great attention in the study of the first few largest eigenvalues and their associated eigenvectors, which are closely related to the principal component analysis (PCA) [28].

Sample covariance matrix [43] plays a central role in the context of modern statistical learning theory, where the data matrix $Y$ is assumed to be $Y = A^{1/2}X$ for some positive definite population covariance matrix $A$, and $X = (x_{ij})$ is the main random source where $x_{ij}$’s are i.i.d. centered random variables. An extension of the sample covariance matrix is the separable covariance matrix [20, 35, 41], where the data matrix is $Y = A^{1/2}XB^{1/2}$ with another positive definite matrix $B$. In spatiotemporal data analysis, $A$ and $B$ are respectively the spatial and temporal covariance matrices.

Even though the assumption that $X$ has i.i.d. entries are popular and useful in the literature, it prohibits the applications of other types of random matrices. An important example is the Haar distributed random matrices. We mention that the Haar distributed random matrices have been used in the study of high
dimensional statistics and machine learning theory, for instance, see [22, 31, 34, 42]. In this paper, we consider that \( X = U \) is either an \( N \times N \) random Haar unitary or orthogonal matrix and assume that

\[
Y = A^{1/2} U B^{1/2}.
\]  

(1.1)

In this sense, we introduce a new class of separable random matrices. The data matrix (1.1) has appeared in the study of high dimensional data analysis, for instance, data acquisition and sensor fusion [19] and matrix denoising [14, 15]. We remark that when \( X = U \) is a Haar random matrix, all the discussions in [20, 41] will be invalid since the independence structure of \( X \) is violated.

The study of the singular values of \( Y \) is closely related to the free convolution of the empirical spectral distributions (ESDs) of \( A \) and \( B \), denoted as \( \mu_A \) and \( \mu_B \), respectively. More specifically, \( YY^* \) has the same eigenvalues with the product of random matrices in generic position, i.e.,

\[
H = AUBU^* ,
\]  

(1.2)

whose global law, i.e., the ESD of \( H \), has been identified in the literature of free probability theory. In the influential work [39], Voiculescu studied the limiting spectral distribution of the eigenvalues of \( H \) and showed that it is given by the free multiplicative convolution of \( \mu_A \) and \( \mu_B \), denoted as \( \mu_A \boxtimes \mu_B \); see Definition 2.7 for a precise definition. More recently, in [27], the author investigated the behavior of \( \mu_A \boxtimes \mu_B \) by analyzing a system of deterministic equations, known as subordination equations, that define the free convolution; see (3.1) for details. They also proved that under certain regularity assumptions, the density of \( \mu_A \boxtimes \mu_B \) has a regular square root behavior near the edge of its support.

In this paper, we first study the singular values and vectors of \( Y \) of the form (1.1), assuming that \( A \) and \( B \) are some positive definitive matrices whose ESDs behave regularly; see Assumptions 2.2 and 2.3 for more details. We establish the local laws for \( Y \), prove the results of eigenvalue rigidity and eigenvector delocalization. Motivated by the applications in statistics, we focus our discussion both near and far away from the rightmost edge of the support of \( \mu_A \boxtimes \mu_B \). Then we study a deformation of \( Y \) in the sense that a few spikes (i.e., eigenvalues that detached from the bulks of the spectrums) are added to \( A \) or \( B \), i.e., the spiked or deformed invariant model [9]

\[
\widehat{Y} = \widehat{A}^{1/2} U \widehat{B}^{1/2},
\]  

(1.3)

where \( \widehat{A} \) and \( \widehat{B} \) are finite-rank perturbations of \( A \) and \( B \), respectively; see equation (2.21) and Section 2.3 for more precise definitions. For (1.3), we study the first order convergent limits and rates for the outlier eigenvalues and their associated generalized components in very general settings. Our results can be used for statistical estimation for the models (1.1) and (1.3) and provide insights for statistical learning problems involving Haar random matrices.

We point out that the addition of random matrices in generic position, i.e., \( A+UBU^* \), has been studied in a series of papers [3, 4, 5, 6, 7]. As mentioned in [7], the results demonstrate that the Haar randomness in the additive model has an analogous behavior to the Wigner matrices [24] in the sense of the strong concentration of the eigenvalues and eigenvectors. In this spirit, the Haar randomness in our multiplicative models, both (1.1) and (1.3), have a similar behavior as the separable covariance matrices in [41, 20]. However, on the technical level, as we mentioned earlier, the technique developed in [41, 20] cannot be applied as the main source of randomness is Haar matrix. To address this issue, we use the technique of partial randomness decomposition (c.f. (6.1)) as developed in [3, 4, 5, 6, 7]. It is the counterpart of the Schur complement in [41, 20]. In this sense, our current paper is on the research line of studying random matrices whose main random resources are Haar matrices. We expect this would be welcome results to researchers who are interested in the intersection of free probability and random matrices.

The core of our proof is to conduct a sophisticated analysis on the subordination functions and their random equivalents in terms of the resolvents of \( H \). Inspired by the arguments in [27], in the current paper, instead of using the subordination functions originally introduced in [38], we use their variants in terms of the M-transform (c.f. Definition 2.1). One advantage of using the M-transform is that it provides us a simple way to write down the approximate subordination functions (c.f. Definition 6.1) in terms of \( (H-z)^{-1}, z \in \mathbb{C}_+ \),
which in turn offers the deterministic limits of the resolvents in terms of the subordination functions (c.f. (2.16) and (2.18)). Similar to the discussion of the additive model in [7], for the rigorous proof, we need to conduct a stability analysis for the system of the subordination equations and control the errors between the subordination functions and their approximates. The stability analysis utilize the Kantorovich’s theorem and the regularity behavior of the subordination functions. To control the error, we first use the partial randomness decomposition to explore some hidden relations, for instance, the error expression between the subordination functions and their approximates in terms of the resolvents; see (6.11) and (6.14) for an illustration. Then we use the device of integration by parts to start the recursive estimates. As mentioned in [7], the weights in the fluctuation averaging mechanisms needed to be properly chosen. In our case, these weights (c.f. (7.34)) can be easily constructed using the hidden identities obtained earlier. We emphasize that in the current paper, we basically follow the proof idea in [7] for the additive model. However, as pointed out in [24], there exist many differences and extra technical difficulties for our multiplicative model, for instance, the counterpart of the M-transform in [7] is simply the negative reciprocal Stieltjes transform which makes their calculation easier than us. Compared to [7], we also establish the bounds for the off-diagonal entries of the resolvents by using the recursive estimate procedure. Such controls are only proved in [5] for the additive model in the bulk case using a more complicated discussion. Moreover, we also establish the local laws for the spectral parameter far away from the edge and all the way down to the real axis. The counterpart for the additive model is only proved for the averaged local law in [7].

Once we have established the local law outside the bulk spectrum, we can use it to study the model (1.3). It is notable that in [9], the authors have studied the convergent limits of the eigenvalues and eigenvectors of \( \hat{Y} Y^* \) to some extent for the Haar unitary matrices under stronger assumptions; see (2.23), (2.24) and Remark 2.24. In the current paper, we will greatly improve the results of [7] by, on one hand, establishing their convergent rates which we believe are optimal up to some \( N^k \) factor for some small \( \epsilon > 0 \), and on the other hand, considering more general assumptions; see Assumption 2.27.

We mention that our results can be used for statistical estimation of the models (1.1) and (1.3). For instance, our local laws, Theorems 2.12 and 2.16 can be used to estimate the subordination functions, which in general are difficult to calculate. Moreover, Theorem 2.20 can be used to estimate the values of the spikes once we have obtained approximations for the subordination functions, and Theorem 2.23 can be used to estimate the number of spikes. All these results can provide useful insights for random matrices involving Haar randomness in statistical learning theory. These will be discussed in Section 2.4. Finally, we believe that both the results and techniques established in this paper can be employed to study other problems. For instance, the Tracy-Widom distribution for the largest eigenvalue of \( H \) in (1.2) and the second order asymptotics of the outlying eigenvalues and eigenvectors for \( \hat{Y} Y^* \) in (1.3). We will pursue these topics in future works.

The rest of the paper is organized as follows. In Section 2, we introduce the necessary notations and state the main results: Section 2.1 is devoted to introducing the notations and assumptions, Section 2.2 is used to state the main results on the local laws, eigenvalue rigidity and eigenvector delocalization, Section 2.3 is devoted to the explanation of the results of the spiked model (1.3) and some discussions on the statistical applications are put in Section 2.4. In Section 2.5 we prove the properties of the free multiplicative convolution and the subordination functions. Section 2.6 is devoted to the proof of the results regarding the spiked model (1.3) and Section 2.7 proves the eigenvalue rigidity and eigenvector delocalization. The other sections are devoted to the proof of the local laws: Section 9 proves the local laws for fixed spectral parameter, Section 10 proves the results of fluctuation averaging and finally, Section 11 establishes the local laws. We put additional technical results in the appendices: Some derivative formulas which will be used through the integration by parts are collected in Appendix A; some auxiliary lemmas regarding large deviation estimates and stability analysis are provided in Appendix B; the continuity arguments for generalizing the domain of spectral parameter are offered in Appendix C and some additional lemmas are proved in Appendix D.

**Conventions.** For \( M, N \in \mathbb{N} \), we denote the set \( \{ k \in \mathbb{N} : M \leq k \leq N \} \) by \([M, N]\). For \( N \in \mathbb{N} \) and \( i \in [1, N] \), we denote by \( e_i^{(N)} \) the \((N \times 1)\) column vector with \((e_i)_j = \delta_{ij}\). We will often omit the superscript \( N \) to write \( e_i^{(N)} \equiv e_i \) as long as there is no confusion. We use \( I \) for the identity matrix of any dimension.
without causing any confusion. We use \( \mathbb{1} \) for the indicator function. For \( N \)-dependent nonnegative numbers \( a_N \) and \( b_N \), we write \( a_N \sim b_N \) if there exists a constant \( C > 1 \) such that \( C^{-1}b_N \leq |a_N| \leq Cb_N \) for all sufficiently large \( N \). Also we write \( a_N = O(b_N) \) or \( a_N \ll b_N \) if \( a_N \leq Cb_N \) for some constant \( C > 0 \) and all sufficiently large \( N \). Finally, we write \( a_N = o(b_N) \) if for all small \( \epsilon > 0 \) there exists \( N_0 \) such that \( a_N \leq \epsilon b_N \) for all \( N \geq N_0 \). We write \( a_N \approx b_N \) if \( a_N = O(b_N) \). Let \( g = (g_1, \ldots, g_N) \) be an \( N \)-dimensional real or complex Gaussian vector. We write \( g \sim N_{R}(0, \sigma^2 I_N) \) if \( g_1, \ldots, g_N \) are i.i.d. \( N(0, \sigma^2) \) random variables; and we write \( g \sim N_{C}(0, \sigma^2 I_N) \) if \( g_1, \ldots, g_N \) are i.i.d. \( N_{C}(0, \sigma^2) \) variables, where \( g_i \sim N_{C}(0, \sigma^2) \) means that \( \text{Re} \ g_i \) and \( \text{Im} \ g_i \) are independent \( N(0, \frac{\sigma^2}{2}) \) random variables. We use \( \mathbb{C}_+ \) to denote the complex upper half plane. For any matrix \( A \), we denote its operator norm by \( \|A\| \) and its Hilbert–Schmidt norm as \( \|A\|_{\text{HS}} \). For a vector \( \nu \), we use \( \|\nu\| \) for its \( \ell_2 \) norm. Moreover, for any complex value \( a \), we use \( \bar{a} \) for its complex conjugate.

Acknowledgments. The authors would like to thank Zhigang Bao and Ji Oon Lee for many helpful discussions and comments. The first author also wants to thank Hau-Tieng Wu for some discussions on the applications in medical sciences.

2 Main results

2.1 Notations and assumptions

For any \( N \times N \) matrix \( W \), we denote its normalized trace by \( \text{tr} \), i.e.,

\[
\text{tr} W = \frac{1}{N} \sum_{i=1}^{N} W_{ii}. \tag{2.1}
\]

Moreover, its empirical spectral distribution (ESD) is defined as

\[
\mu_W = \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i(W)}.
\]

In the present paper, even if the matrix is not of size \( N \times N \), the trace is always normalized by \( N^{-1} \) unless otherwise specified.

We first introduce our model. Consider two \( N \times N \) real deterministic positive definite matrices

\[
A \equiv A_N = \text{diag}(a_1, \ldots, a_N) \quad \text{and} \quad B \equiv B_N = \text{diag}(b_1, \ldots, b_N),
\]

where the diagonal entries are ordered as \( a_1 \geq a_2 \geq \cdots \geq a_N > 0 \) and \( b_1 \geq b_2 \geq \cdots \geq b_N > 0 \). Let \( U \equiv U_N \) be a random unitary or orthogonal matrix, Haar distributed on the \( N \)-dimensional unitary group \( U(N) \) or orthogonal group \( O(N) \). Define \( \bar{A} := U^* A U \), \( \bar{B} := U B U^* \), and

\[
H := A U B U^*, \quad \mathcal{H} := U^* A U B, \quad \tilde{H} := A^{1/2} \bar{B} A^{1/2}, \quad \text{and} \quad \tilde{\mathcal{H}} := B^{1/2} \tilde{A} B^{1/2}. \tag{2.2}
\]

Note that we only need to consider diagonal matrices \( A \) and \( B \) since \( U \) is a Haar random unitary or orthogonal matrix. Moreover, \( \bar{H} \) and \( \tilde{\mathcal{H}} \) are Hermitian random matrices.

Since \( H, \mathcal{H}, \tilde{H} \) and \( \tilde{\mathcal{H}} \) have the same eigenvalues, in the sequel, we denote the eigenvalues of all of them as \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \) without causing any confusion. Further, we define the empirical spectral distributions (ESD) of the above matrices by

\[
\mu_A \equiv \mu_A^{(N)} := \frac{1}{N} \sum_{i=1}^{N} \delta_{\lambda_i}, \quad \mu_B \equiv \mu_B^{(N)} := \frac{1}{N} \sum_{i=1}^{N} \delta_{\mu_i} \quad \text{and} \quad \mu_H \equiv \mu_H^{(N)} := \frac{1}{N} \sum_{i=1}^{N} \delta_{\lambda_i}.
\]

For \( z \in \mathbb{C}_+ := \{ z \in \mathbb{C} : \text{Im} z > 0 \} \), we define the resolvents of the above random matrices as follows

\[
G(z) := (H - zI)^{-1}, \quad \mathcal{G}(z) := (\mathcal{H} - zI)^{-1}, \quad \tilde{G}(z) := (\tilde{H} - zI)^{-1}, \quad \tilde{\mathcal{G}}(z) := (\tilde{\mathcal{H}} - zI)^{-1}. \tag{2.3}
\]
In the rest of the paper, we usually omit the dependence of \( z \) and simply write \( G, \mathcal{G}, \widetilde{G} \) and \( \widetilde{\mathcal{G}} \). The following transforms will play important roles in the current paper.

**Definition 2.1.** For any probability measure \( \mu \) defined on \( \mathbb{R}_+ \), its Stieltjes transform \( m_\mu \) is defined as

\[
m_\mu(z) := \int \frac{1}{x - z} d\mu(x), \quad \text{for } z \in \mathbb{C} \setminus \mathbb{R}_+.
\]

Moreover, we define the M-transform \( M_\mu \) and L-transform \( L_\mu \) on \( \mathbb{C} \setminus \mathbb{R}_+ \) as

\[
M_\mu(z) := 1 - \left( \int \frac{x}{x - z} d\mu(x) \right)^{-1} = \frac{zm_\mu(z)}{1 + zm_\mu(z)}, \quad L_\mu(z) := \frac{M_\mu(z)}{z}.
\]

Let \( \mu_H(z) \) be the Stieltjes transform of the ESD of \( H \). Since \( H, \mathcal{H}, \widetilde{H}, \widetilde{\mathcal{H}} \) are similar to each other, we have that \( m_H(z) = \text{tr } G = \text{tr } \mathcal{G} = \text{tr } \widetilde{G} = \text{tr } \widetilde{\mathcal{G}} \). Moreover, it is easy to see that

\[
G_{ij} = \sqrt{a_i/a_j} \widetilde{G}_{ij}, \quad \widetilde{G}_{ij} = \sqrt{b_i/b_j} \mathcal{G}_{ij}.
\]

We next introduce the main assumptions of the paper. Throughout the paper, we assume that there exist two \( N \)-independent absolutely continuous probability measures \( \mu_\alpha \) and \( \mu_\beta \) on \((0, \infty)\) satisfying the following conditions.

**Assumption 2.2.** Suppose the following assumptions hold true:

(i). \( \mu_\alpha \) and \( \mu_\beta \) have densities \( \rho_\alpha \) and \( \rho_\beta \), respectively. For the ease of discussion, we assume that both of them have means 1, i.e.,

\[
\int xd\mu_\alpha(x) = \int x\rho_\alpha(x)dx = 1.
\]

(ii). Both \( \rho_\alpha \) and \( \rho_\beta \) have single non-empty intervals as supports, denoted as \([E_\alpha^\alpha, E_\alpha^\beta]\) and \([E_\beta^\alpha, E_\beta^\beta]\), respectively. Here \( E_\alpha^\alpha, E_\alpha^\beta, E_\beta^\alpha \) and \( E_\beta^\beta \) are all positive numbers. Moreover, both of the density functions are strictly positive in the interior of their supports.

(iii). There exist constants \(-1 < t_\alpha^\pm, t_\beta^\pm < 1\) and \( C > 1 \) such that

\[
C^{-1} \leq \frac{\rho_\alpha(x)}{(x - E_\alpha^\alpha)^{t_\alpha^\pm}(E_\alpha^\alpha - x)^{t_\alpha^\pm}} \leq C, \quad \forall x \in [E_\alpha^\alpha, E_\alpha^\beta],
\]

\[
C^{-1} \leq \frac{\rho_\beta(x)}{(x - E_\beta^\beta)^{t_\beta^\pm}(E_\beta^\beta - x)^{t_\beta^\pm}} \leq C, \quad \forall x \in [E_\beta^\alpha, E_\beta^\beta].
\]

**Remark 2.3.** The assumption (i) is introduced for technical simplicity and it can be removed easily; see Remark 3.2 of [27] for details. Moreover, the assumption (iii) is introduced to guarantee the square root behavior near the edges of the free multiplicative convolution of \( \mu_\alpha \) and \( \mu_\beta \). When this condition fails, the behavior of \( \mu_\alpha \boxplus \mu_\beta \) near the edge can be very different from our current discussion; see [30] for more details.

The second assumption, Assumption 2.4 ensures that \( \mu_A \) and \( \mu_B \) are close to \( \mu_\alpha \) and \( \mu_\beta \), respectively. Specifically, it demonstrates that the convergence rates of \( \mu_A \) and \( \mu_B \) to \( \mu_\alpha \) and \( \mu_\beta \) are bounded by an order of \( N^{-1} \), so that their fluctuations do not dominate that of \( \mu_H \).

**Assumption 2.4.** Suppose the following assumptions hold true:

(iv). For the Levy distance \( \mathcal{L}(\cdot, \cdot) \), we have that for any small constant \( \epsilon > 0 \)

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

when \( N \) is sufficiently large.
Moreover, we conclude from [27, Theorem 3.1] that, if (i) and (ii) in Assumption 2.2 hold, then

\[
\text{absolutely continuous and supported on a single non-empty compact interval on } (0, \infty),
\]

when \( N \) is sufficiently large.

Remark 2.5. We remark that we will consistently use \( \epsilon \) as a sufficiently small constant for the rest of the paper. The assumption (v) assures that both of the upper edges of \( \mu_A \) and \( \mu_B \) are bounded.

As proved by Voiculescu in [38, 39], under Assumptions 2.2 and 2.4, \( \mu_H \) converges weakly to a deterministic measure, denoted as \( \mu_\alpha \boxtimes \mu_\beta \). It is called the free multiplicative convolution of \( \mu_\alpha \) and \( \mu_\beta \). In the present paper, we use the M-transform in (2.4) to define the free multiplicative convolution.

Lemma 2.6 (Proposition 2.5 of [27]). For Borel probability measures \( \mu_\alpha \) and \( \mu_\beta \) on \( \mathbb{R}_+ \), there exist unique analytic functions \( \Omega_\alpha, \Omega_\beta : \mathbb{C} \setminus \mathbb{R}_+ \to \mathbb{C} \setminus \mathbb{R}_+ \) satisfying the following:

1. For all \( z \in \mathbb{C}_+ \), we have

\[
\arg \Omega_\alpha(z) \geq \arg z \quad \text{and} \quad \arg \Omega_\beta(z) \geq \arg z.
\]

2. For all \( z \in \mathbb{C}_+ \),

\[
\lim_{z \to -\infty} \Omega_\alpha(z) = \lim_{z \to -\infty} \Omega_\beta(z) = -\infty.
\]

3. For all \( z \in \mathbb{C} \setminus \mathbb{R}_+ \), we have

\[
z \mu_\alpha(\Omega_\beta(z)) = z \mu_\beta(\Omega_\alpha(z)) = \Omega_\alpha(z) \Omega_\beta(z).
\]

With Lemma 2.6, we now define the free multiplicative convolution.

Definition 2.7. Denote the analytic function \( M : \mathbb{C} \setminus \mathbb{R}_+ \to \mathbb{C} \setminus \mathbb{R}_+ \) by

\[
M(z) := M_{\mu_\alpha}(\Omega_\beta(z)) = M_{\mu_\beta}(\Omega_\alpha(z)).
\]

The free multiplicative convolution of \( \mu_\alpha \) and \( \mu_\beta \) is defined as the unique probability measure \( \mu \), denoted as \( \mu = \mu_\alpha \boxtimes \mu_\beta \) such that (2.7) holds for all \( z \in \mathbb{C} \setminus \mathbb{R}_+ \). In this sense, \( M(z) = M_{\mu_\alpha \boxtimes \mu_\beta}(z) \) is the M-transform of \( \mu_\alpha \boxtimes \mu_\beta \). Furthermore, the analytic functions \( \Omega_\alpha \) and \( \Omega_\beta \) are referred to as the subordination functions. Similarly, we define \( \Omega_A \) and \( \Omega_B \) by replacing \( (\alpha, \beta) \) with \( (A, B) \) in Lemma 2.6 and define \( \mu_A \boxtimes \mu_B \) so that

\[
M_{\mu_A}(\Omega_B(z)) = M_{\mu_B}(\Omega_A(z)) = M_{\mu_A \boxtimes \mu_B}(z) \text{ for all } z \in \mathbb{C} \setminus \mathbb{R}_+.
\]

Note that a straightforward consequence of (2.8) and the definition of \( M_{\mu}(z) \) is the following identity

\[
\int \frac{x}{x-z} d(\mu_\alpha \boxtimes \mu_\beta)(x) = zm_{\mu_\alpha \boxtimes \mu_\beta}(z) + 1 = \Omega_\beta(z) m_{\mu_\alpha}(\Omega_\beta(z)) + 1 = \int \frac{x}{x-\Omega_\beta(z)} d\mu_\alpha(x).
\]

Remark 2.8. Since all of \( \mu_\alpha, \mu_\beta, \mu_A \), and \( \mu_B \) are compactly supported in \((0, \infty)\), similar results hold for \( \mu_\alpha \boxtimes \mu_\beta \) and \( \mu_A \boxtimes \mu_B \). Specifically, we have [40] Remark 3.6.2. (iii)]

\[
supp \mu_\alpha \boxtimes \mu_\beta \subset [E^\alpha_-, E^\beta_+], \quad supp \mu_A \boxtimes \mu_B \subset [a_N b_N, a_1 b_1].
\]

Moreover, we conclude from [27, Theorem 3.1] that, if (i) and (ii) in Assumption 2.2 hold, then \( \mu_\alpha \boxtimes \mu_\beta \) is absolutely continuous and supported on a single non-empty compact interval on \((0, \infty)\), denoted as \([E_-, E_+]\), i.e.,

\[
E_- := \inf supp(\mu_\alpha \boxtimes \mu_\beta), \quad E_+ := \sup supp(\mu_\alpha \boxtimes \mu_\beta).
\]

Moreover, let the density of \( \mu_\alpha \boxtimes \mu_\beta \) be \( \rho \). For small constant \( \tau > 0 \), with (iii) of Assumption 2.2 we have

\[
\rho(x) \sim \sqrt{E_+ - x}, \quad x \in [E_+ - \tau, E_+].
\]

\(^1\)Note that the support of a non-negative measure is defined to be the largest (closed) subset such that for which every open neighbourhood of every point of the set has positive measure. The notation does not take isolated singletons into consideration.
Remark 2.9. As we will see later in Lemma 5.4, the subordination functions of $\Omega_\alpha$ and $\Omega_\beta$ will also have square root behaviors near the edges. The regularity behavior is assured by the fact that the subordination functions $\Omega_\alpha(\beta)$ are well separated from the supports of $\mu_\beta(\mu_\alpha)$, i.e., (ii) of Lemma 3.2. In fact, from the proof of Proposition 5.6 of [27], we see that the assumption (iii) in Assumption 2.2 implies this stability result.

Remark 2.10. It is known from [8, 27] that Assumption 2.2 ensures that the subordination functions $\Omega_\alpha|_{C_+}$ and $\Omega_\beta|_{C_+}$ can be extended continuously to the real line. Till the end of the paper, we will write $\Omega_\alpha(x)$ or $\Omega_\beta(x)$ for $x \in \mathbb{R}$ to denote the values of the continuous extensions. In particular, $\Omega_\alpha(x)$ and $\Omega_\beta(x)$ always have nonnegative imaginary parts for all $x \in \mathbb{R}$.

### 2.2 Local laws for free multiplication of random matrices

In this section, we state the results of the local laws. We will need the following notation of stochastic domination to state our main results. It was first introduced in [23] and subsequently used in many works on random matrix theory. It simplifies the presentation of the results and the proofs by systematizing statements of the form “$X_N$ is bounded by $Y_N$ with high probability up to a small power of $N$”.

**Definition 2.11.** For two sequences of random variables $\{X_N\}_{N \in \mathbb{N}}$ and $\{Y_N\}_{N \in \mathbb{N}}$, we say that $X_N$ is stochastically dominated by $Y_N$, written as $X_N \prec Y_N$ or $X_N = O_+(Y_N)$, if for all (small) $\epsilon > 0$ and (large) $D > 0$, we have

$$\mathbb{P} \left[ |X_N| \geq N^\epsilon |Y_N| \right] \leq N^{-D},$$

for sufficiently large $N \geq N_0(\epsilon, D)$. If $X_N(v)$ and $Y_N(v)$ depend on some common parameter $v$, we say $X_N \prec Y_N$ uniformly in $v$ if the threshold $N_0(\epsilon, D)$ can be chosen independent of the parameter $v$.

Moreover, we say an event $\Xi$ holds with high probability if for any constant $D > 0$, $\mathbb{P}(\Xi) \geq 1 - N^{-D}$ for large enough $N$.

We first provide some notations. For any spectral parameter $z = E + i\eta \in \mathbb{C}_+$, we define

$$\kappa \equiv \kappa(z) := |E - E_+|,$$

where $E_+$ is the rightmost edge of $\mu_\alpha \boxtimes \mu_\beta$ given in (2.12). For given constants $0 \leq a \leq b$ and small constant $0 < \tau < \min\left\{ \frac{E_+ - E_-}{2}, 1 \right\}$, we define the following set of spectral parameter $z$ by

$$D_\tau(a, b) := \{ z = E + i\eta \in \mathbb{C}_+ : E_+ - \tau \leq E \leq \tau^{-1}, a \leq \eta \leq b \}. \quad (2.14)$$

Further, for any small positive $\gamma > 0$, we let

$$\eta_L \equiv \eta_L(\gamma) := N^{-1+\gamma}, \quad (2.15)$$

and let $\eta_U > 1$ be a large $N$-independent constant. The following theorem establishes the local laws for the matrices $H, \tilde{H}, \mathcal{H}$ and $\tilde{\mathcal{H}}$ near the upper edge $E_+$. Analogous results can be obtained for the lower edge $E_-$. 

**Theorem 2.12** (Local laws near the edge). Suppose Assumptions 2.2 and 2.4 hold. Let $\tau$ and $\gamma$ be fixed small positive constants. For any deterministic vector $v = (v_1, \cdots, v_N) \in \mathbb{C}$ such that $\|v\|_\infty \leq 1$, we have:

1. For the matrix $H$ and its resolvent $G(z)$, we have

$$\left| \frac{1}{N} \sum_{i=1}^{N} v_i \left( zG_{ii}(z) + 1 - \frac{a_i}{a_i - \Omega_{ii}(z)} \right) \right| \lesssim \frac{1}{N\eta}, \quad (2.16)$$

uniformly in $z \in D_\tau(\eta_L, \eta_U)$ with $\eta_L$ defined in (2.15) and any fixed constant $\eta_U$. In particular, we have

$$|m_H(z) - m_{\mu_\alpha \boxtimes \mu_\beta}(z)| \lesssim \frac{1}{N\eta},$$
Moreover, for the off-diagonal entries, we have
\[ \max_{i \neq j} |G_{ij}(z)| \prec \frac{1}{\sqrt{N\eta}}. \] (2.17)

Similar results hold true by simply replacing \( H \) and \( G(z) \) with \( \tilde{H} \) and \( \tilde{G}(z) \), respectively.

(2) For the matrix \( H \) and its resolvent \( G(z) \), we have \[ \left| \frac{1}{N} \sum_{i=1}^{N} v_i \left( zG_{ii}(z) + 1 - \frac{b_i}{b_i - \Omega_A(z)} \right) \right| \prec \frac{1}{N\eta}. \] (2.18)

In particular, we have
\[ |m_H(z) - m_{\mu_A \boxtimes \mu_B}(z)| \prec \frac{1}{N\eta}, \]
uniformly in \( z \in D_\tau(\eta_L, \eta_U) \). Moreover, for the off-diagonal entries, we have
\[ \max_{i \neq j} |G_{ij}(z)| \prec \frac{1}{\sqrt{N\eta}}. \] (2.19)

Similar results hold true by simply replacing \( H \) and \( G(z) \) with \( \tilde{H} \) and \( \tilde{G}(z) \), respectively.

Remark 2.13. We remark that in Theorem 2.5 of [7], which is the counterpart for the additive model, only the local laws for the diagonal elements of the resolvents are established. We also consider the off-diagonal entries and prove that they are small. Moreover, in Proposition 4.1, we see that all the bounds in the above theorem can be replaced by
\[ \sqrt{\text{Im} m_{\mu_A \boxtimes \mu_B}(z)/N \eta + 1/N\eta}, \]
which matches the typical forms of the bounds of local laws in Random Matrix Theory literature, for instance, see the monograph [24]. We keep the current form to highlight both the similarity and differences between our multiplicative model and the additive model in [7].

Denote \( \gamma_j \) as the \( j \)-th \( N \)-quantile (or typical location) of \( \mu_\alpha \boxtimes \mu_\beta \) such that
\[ \int_{\gamma_j}^{\infty} d\mu_\alpha \boxtimes \mu_\beta(x) = \frac{j}{N}. \]
Similarly, we denote \( \gamma_j^* \) to be the \( j \)-th \( N \)-quantile of \( \mu_A \boxtimes \mu_B \). Recall that \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \) are the eigenvalues of \( AUBU^* \). We next state two important consequences of our local laws.

**Theorem 2.14** (Spectral rigidity near the upper edge). Suppose Assumptions 2.2 and 2.4 hold true. For any small constant \( 0 < c < 1/2 \), we have that for all \( 1 \leq i \leq cN \),
\[ |\lambda_i - \gamma_i^*| \prec i^{-1/3} N^{-2/3}. \]

Moreover, the same conclusion holds if \( \gamma_i^* \) is replaced with \( \gamma_i \).

Consider (1.1) that \( Y = A^{1/2}UB^{1/2} \). Denote its singular value decomposition (SVD)
\[ Y = \sum_{k=1}^{N} \sqrt{\lambda_k} u_k v_k^*, \]
where \( \{u_k\} \) and \( \{v_k\} \) are the left and right singular vectors of \( Y \), respectively.
Theorem 2.15 (Delocalization of the singular vectors). Suppose Assumptions 2.2 and 2.4 hold true. For any small constant $0 < c < 1/2$, we have that for all $1 \leq k \leq cN$, 

$$
\max_i |u_k(i)|^2 + \max_{\mu} |v_k(\mu)|^2 < \frac{1}{N}.
$$

Finally, outside the spectrum, we have stronger control all the way down to the real axis. We record such results in the following theorem, which will be the key technical input for the proof of the spiked model. For the parameter $\tau$ in (2.14), we denote the spectral domain of the parameter as

$$
\mathcal{D}_\tau(\eta_U) := \{ z = E + \eta \in \mathbb{C}_+ : E_+ + N^{-2/3+\tau} \leq E \leq \tau^{-1}, 0 \leq \eta \leq \eta_U \}.
$$

Theorem 2.16 (Local laws far away from the spectrum). Suppose Assumptions 2.2 and 2.4 hold. Let $\tau$ be a fixed small positive constant. We have the following hold true uniformly in $z \in \mathcal{D}_\tau(\eta_U)$: 

(1). For the matrix $H$, we have 

$$
|m_H(z) - m_{\mu_A}(z)| < \frac{1}{N(\kappa + \eta)},
$$

Similar results hold if we replace $H$ with $\mathcal{H}, \mathcal{H}$ and $\mathcal{H}$. 

(2). For the resolvent $G$, we have 

$$
\max_i \left| zG_{ii}(z) + 1 - \frac{a_i}{a_i - \Omega_B(z)} \right| < N^{-1/2}(\kappa + \eta)^{-1/4},
$$

and 

$$
\max_{i,j} |G_{ij}(z)| < N^{-1/2}(\kappa + \eta)^{-1/4}.
$$

Similar results hold for $G, \tilde{G}$ and $\tilde{G}$. 

2.3 Spiked invariant model 

In this section, we employ the local laws obtained in Section 2.2 to study the eigenvalues and eigenvectors of spiked model [18] and improve the results obtained in [9] Section 2.2, which only concerns the Haar unitary random matrices. To add a few spikes, we follow the setup of [18, 20] and assume that there exist some fixed integers $r$ and $s$ with two sequences of positive numbers $\{d_i^r\}_{i \leq r}$ and $\{d_j^s\}_{j \leq s}$ such that $\bar{A} = \text{diag} \{\hat{a}_1, \cdots, \hat{a}_N\}$ and $\bar{B} = \text{diag} \{\hat{b}_1, \cdots, \hat{b}_N\}$, where

\[
\hat{a}_k = \begin{cases} 
    a_k(1 + d_k^r), & 1 \leq k \leq r \\
    a_k, & k \geq r + 1,
\end{cases} \quad \hat{b}_k = \begin{cases} 
    b_k(1 + d_k^s), & 1 \leq k \leq s \\
    b_k, & k \geq s + 1.
\end{cases}
\] (2.21)

Without loss of generality, we assume that $\hat{a}_1 \geq \hat{a}_2 \cdots \geq \hat{a}_N$ and $\hat{b}_1 \geq \hat{b}_2 \cdots \geq \hat{b}_N$. In the current paper, we assume that all the $d_k^r$'s and $d_k^s$'s are bounded.

We will investigate the behavior of the singular values and vectors of $\tilde{Y} = \bar{A}^{1/2}U\tilde{B}^{1/2}$. We denote $\tilde{Q}_1 = \tilde{Y}Y^*$ and $\tilde{Q}_2 = \tilde{Y}^*\tilde{Y}$. Denote the eigenvalues of $\tilde{Q}_1$ and $\tilde{Q}_2$ as $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots \geq \tilde{\lambda}_N$. Recall that $\Omega_A(\cdot)$ and $\Omega_B(\cdot)$ are the subordination functions associated with $\mu_A$ and $\mu_B$ and Remark 2.10. We will see that if a spike $\hat{a}_i, 1 \leq i \leq r$ or $\hat{b}_j, 1 \leq j \leq s$, causes an outlier eigenvalue, if

$$
\hat{a}_i > \Omega_B(E_+), \text{ or } \hat{b}_j > \Omega_A(E_+).
$$

We first introduce the following assumption.
Assumption 2.17. We assume that (2.14) holds for all $1 \leq i \leq r$ and $1 \leq j \leq s$. Moreover, we define the integers $0 \leq i \leq r^+$ and $0 \leq j \leq s^+$ such that

$$\hat{a}_i \geq \Omega_B(E_+) + N^{-1/3} \text{ if and only if } 1 \leq i \leq r^+, \tag{2.21}$$

and

$$\hat{b}_j \geq \Omega_A(E_+) + N^{-1/3} \text{ if and only if } 1 \leq j \leq s^+. \tag{2.22}$$

The lower bound $N^{-1/3}$ is chosen for definiteness, and it can be replaced with any $N$-dependent parameter that is of the same order.

Remark 2.18. A spike $\hat{a}_i$ or $\hat{b}_j$ that does not satisfy the assumptions in Assumption 2.17 will cause an outlying eigenvalue that lies within an $O_<(N^{-2/3})$ neighborhood of the edge $E_+$. In this sense, it will be hard to detect such spikes as we have seen from Theorem 2.20 the eigenvalues of $Q_1 = YY^*$ near the edge also have a fluctuation of order $N^{-2/3}$. Assumption 2.17 simply choose the “real” spikes. In the statistical literature, this is referred to as the supercritical regime and a reliable detection of the spikes is only available in this regime. We refer the readers to [2] [13] [20] [35] for more detailed discussion.

To state the results of the outlier and extremal non-outlier eigenvalues, we introduce the following definition following [20] Definition 3.5. Since both $\Omega_A(x)$ and $\Omega_B(x)$ are real and monotone increasing for $x \geq E_+$, we can then denote $\Omega_A^{-1}(\cdot)$ and $\Omega_B^{-1}(\cdot)$ the inverse functions of $\Omega_A$ and $\Omega_B$ on $[\Omega_A(E_+), \infty)$ and $[\Omega_B(E_+), \infty)$, respectively.

Definition 2.19. We define the labelling functions $\pi_a : \{1, \ldots, N\} \rightarrow N$ and $\pi_b : \{1, \ldots, N\} \rightarrow N$ as follows. For any $1 \leq i \leq r$, we assign to it a label $\pi_a(i) \in \{1, \ldots, r + s\}$ if $\Omega_B^{-1}(\hat{a}_i)$ is the $\pi_a(i)$-th largest element in $\{\Omega_B^{-1}(\hat{a}_i)\}_{i=1}^{r+s} \cup \{\Omega_A^{-1}(\hat{b}_j)\}_{j=1}^{s}$. We also assign to any $1 \leq j \leq s$ a label $\pi_b(j) \in \{1, \ldots, r + s\}$ in a similar way. Moreover, we define $\pi_a(i) = i + s$ if $i > r$ and $\pi_b(j) = j + r$ if $j > s$. We define the following sets of outlier indices:

$$\mathcal{O} := \{\pi_a(i) : 1 \leq i \leq r\} \cup \{\pi_b(j) : 1 \leq j \leq s\},$$

and

$$\mathcal{O}^+ := \{\pi_a(i) : 1 \leq i \leq r^+\} \cup \{\pi_b(j) : 1 \leq j \leq s^+\}. \tag{2.24}$$

Theorem 2.20 (Eigenvalue statistics). Suppose Assumptions 2.4 2.7 and 2.17 hold. Then we have

$$\left|\hat{\lambda}_{\pi_a(i)} - \Omega_B^{-1}(\hat{a}_i)\right| \sim N^{-1/2}(\hat{a}_i - \Omega_B(E_+))^{1/2}, 1 \leq i \leq r^+, \tag{2.23}$$

and

$$\left|\hat{\lambda}_{\pi_b(j)} - \Omega_A^{-1}(\hat{b}_j)\right| \sim N^{-1/2}(\hat{b}_j - \Omega_A(E_+))^{1/2}, 1 \leq j \leq s^+. \tag{2.24}$$

Moreover, for any fixed integer $\varpi > r + s$, we have

$$\left|\hat{\lambda}_i - E_+\right| \sim N^{-2/3}, \text{ for } i \notin \mathcal{O}^+ \text{ and } i \leq \varpi. \tag{2.22}$$

Remark 2.21. We remark that the convergent limits of the outlier eigenvalues have been obtained under stronger assumptions in [20] Section 2.2] for the spiked unitarily invariant model, where the conditions in Assumption 2.17 are strengthened to

$$\hat{a}_i \geq \Omega_B(E_+) + \zeta \text{ if and only if } 1 \leq i \leq r^+, \tag{2.23}$$

and

$$\hat{b}_j \geq \Omega_A(E_+) + \zeta \text{ if and only if } 1 \leq j \leq s^+, \tag{2.24}$$
where \( \zeta > 0 \) is some fixed constant. We extend the results in [9] in two aspects: considering the more general Assumption 2.17 and establishing their convergent rates. We believe that Assumption 2.17 is the most general assumption possible for the existence of the outliers, and the convergent rates obtained here are optimal up to some \( N' \) factor, where \( \epsilon > 0 \) is some small constant. These results will be used for the discussion of statistical applications in Section 2.4. Finally, we mention that we can follow the discussion in [13, Theorem 2.7] or [20, Theorem 3.7] to show that the non-outlier eigenvalues of \( \hat{Q}_1 \) will be close to those of \( Q_1 \), which is called the eigenvalue sticking property. We will pursue this direction somewhere else.

Then we introduce the results regarding the singular vectors. Specifically, we will state our results under the so-called non-overlapping condition, which was first introduced in [13]. In what follows, we consider an index set \( S \) such that

\[
S \subset O^+.
\]

For convenience, we introduce the following notations. For \( 1 \leq i_1 \leq r^+, 1 \leq i_2 \leq N \) and \( 1 \leq j \leq N \), we define

\[
\delta^a_{\pi_a(i_1), \pi_a(i_2)} := |\tilde{a}_{i_1} - \tilde{a}_{i_2}|, \quad \delta^a_{\pi_a(i_1), \pi_b(j)} := |\tilde{b}_j - \Omega_A(\Omega_B^{-1}(\tilde{a}_{i_1}))|.
\]

Similarly, for \( 1 \leq j_1 \leq s^+, 1 \leq j_2 \leq N \) and \( 1 \leq i \leq N \), we define

\[
\delta^b_{\pi_a(j_1), \pi_a(i)} := |\tilde{a}_i - \Omega_B(\Omega_A^{-1}(\tilde{b}_{j_1}))|, \quad \delta^b_{\pi_a(j_1), \pi_b(j_2)} := |\tilde{b}_{j_1} - \tilde{b}_{j_2}|.
\]

Further if \( a \in S \), we define

\[
\delta_a(S) := \bigg\{ \min_{k: \pi_a(k) \notin S} \delta^a_{\pi_a(k), \pi_a(i)} \bigg\} \wedge \bigg\{ \min_{j: \pi_a(j) \notin S} \delta^a_{\pi_a(j), \pi_a(i)} \bigg\}, \quad \text{if } a = \pi_a(i) \in S \;
\]

\[
\text{if } a = \pi_b(j) \in S ;
\]

and if \( a \notin S \), then we define

\[
\delta_a(S) := \bigg\{ \min_{k: \pi_a(k) \in S} \delta^a_{\pi_a(k), a} \bigg\} \wedge \bigg\{ \min_{j: \pi_b(j) \in S} \delta^b_{\pi_b(j), a} \bigg\}.
\]

We are now ready to impose some assumptions.

**Assumption 2.22.** For some fixed small constants \( \tau_1, \tau_2 > 0 \), we assume that for \( \pi_a(i) \in S \) and \( \pi_b(j) \in S \),

\[
\tilde{a}_i - \Omega_B(E_+) \geq N^{-1/3 + \tau_1}, \quad \tilde{b}_j - \Omega_A(E_+) \geq N^{-1/3 + \tau_1}.
\]

Moreover, we assume that

\[
\delta_{\pi_a(i)}(S) \geq N^{-1/2 + \tau_2} (\tilde{a}_i - \Omega_B(E_+))^{-1/2}, \quad \delta_{\pi_b(j)}(S) \geq N^{-1/2 + \tau_2} (\tilde{b}_j - \Omega_A(E_+))^{-1/2}.
\]

In the above assumption, (2.29) implies that for \( \pi_a(i) \in S \) and \( \pi_b(j) \in S \), \( \{\tilde{a}_i\} \) and \( \{\tilde{b}_j\} \) are the "real" spikes, whereas (2.30) guarantees a phenomenon of cone concentration; see Remark 2.25 for more detailed discussion. With the above preparation, we proceed to state our main results regarding the singular vectors. Let the left singular vectors of \( \hat{Y} \) be \( \{\hat{u}_i\} \) and the right singular vectors be \( \{\hat{v}_i\} \). We denote

\[
P_S = \sum_{k \in S} \hat{u}_k \hat{u}_k^*, \quad \text{and } P'_S = \sum_{k \in S} \hat{v}_k \hat{v}_k^*.
\]

**Theorem 2.23** (Eigenvector statistics). Suppose that Assumptions 2.2, 2.4 and 2.22 hold. For the set \( S \) in (2.22) and any given deterministic vectors \( v = (v_1, \cdots, v_N)^* \in \mathbb{C}^N \), we have that for the left singular vectors,

\[
|\langle v, P_S v \rangle - g_a(v, S)| \leq \sum_{i: \pi_a(i) \in S} \frac{|v_i|^2}{\sqrt{N(\tilde{a}_i - \Omega_B(E_+))}} + \sum_{i=1}^N \frac{|v_i|^2}{N \delta_{\pi_a(i)}(S)} + g_a(v, S)^{1/2} \left( \sum_{\pi_a(i) \notin S} \frac{|v_i|^2}{N \delta_{\pi_a(i)}(S)} \right)^{1/2},
\]
where \( g_a(\mathbf{v}, S) \) is defined as
\[
g_a(\mathbf{v}, S) := \sum_{i : \pi_a(i) \in S} \tilde{a}_i \frac{(\Omega_B^{-1})'_{ii}(\tilde{a}_i)}{\Omega_B'(\tilde{a}_i)} |v_i|^2.
\]

Similarly, for the right singular vectors, we have
\[
|\langle \mathbf{v}, \mathcal{P}_S^* \mathbf{v} \rangle - g_b(\mathbf{v}, S)| < \sum_{i : \pi_b(i) \in S} \frac{|v_i|^2}{\sqrt{N(b_j - \Omega_A(E_+))}} + \sum_{j=1}^{N} \frac{|v_j|^2}{N} N \delta_{\pi_a(j)}(S)^{1/2} + g_b(\mathbf{v}, S)^{1/2} \left( \sum_{\pi_b(j) \notin S} \frac{|v_j|^2}{N \delta_{\pi_a(j)}(S)} \right)^{1/2},
\]
where \( g_b(\mathbf{v}, S) \) is defined as
\[
g_b(\mathbf{v}, S) := \sum_{j : \pi_b(j) \in S} \tilde{b}_j \frac{(\Omega_A^{-1})'_{jj}(\tilde{b}_j)}{\Omega_A'(\tilde{b}_j)} |v_j|^2.
\]

**Remark 2.24.** We mention that some partial results of Theorem 2.23 have been obtained in (4) of Theorem 2.5 in [9] under stronger assumptions for the spiked unitarily invariant model. More specifically, by assuming that \( r = 0 \) or \( s = 0 \), and (2.22) and (2.24), they obtained the concentration limit (2.31). We extend the counterparts in [9] by one one hand, stating the results in great generality under Assumption 2.22 and on the other hand, establishing their convergent rates. Moreover, we remark that we can obtain the results of Theorem 2.23 without Assumption 2.22 following the arguments of [20] Sections S.5.2 and S.6. Finally, for the non-outlier singular vectors, we can prove a delocalization result similar to [20, Theorem 3.14]. Both of the extensions will need more dedicate efforts and will be the topics of future study.

**Remark 2.25.** We provide an example for illustration. For simplicity, we consider the non-degenerate case such that all the outliers are well-separated in the sense that we can simply choose \( S = \{\pi_a(i)\} \) or \( S = \{\pi_b(j)\} \). Let \( S = \{\pi_a(i)\} \) and \( \mathbf{v} = \mathbf{e}_i \). Then we obtain from Theorem 2.23 that
\[
|\langle \mathbf{u}_i, \mathbf{e}_i \rangle|^2 = \tilde{a}_i \frac{(\Omega_B^{-1})'_{ii}(\tilde{a}_i)}{\Omega_B'(\tilde{a}_i)} + O_{\mathbb{P}} \left( \frac{1}{\sqrt{N}} + \frac{1}{N \delta_i^2} \right), \quad \delta_i = \delta_{\pi_a(i)}(\pi_a(i)).
\]

It is easy to see that the non-overlapping condition (2.30), together with the estimates (1.24) and (1.29), imply that the error term is much smaller than the first term of the right-hand side of (2.32). In this sense, \( \mathbf{u}_i \) is concentrated on a cone with axis parallel to \( \mathbf{e}_i \).

### 2.4 Remarks on statistical applications

In this section, we discuss how we can apply the results of Theorems 2.12, 2.14, 2.15, 2.20 and 2.23 to study the models [1.1] and [1.3]. In all the discussions of this subsection, we assume that \( r^+ = r \) and \( s^+ = s \), i.e., all the spikes are in the supercritical regime, which is the most interesting regime in statistics.

As we have seen from the above theorems, all the results involve the subordination functions. Even though we know both \( A \) and \( B \), it is generally difficult to calculate these functions, even numerically. In this sense, the results of Theorems 2.12 and 2.16 provide us a simple way to approximate the subordination functions and estimate the quantities in which we are interested. For example, we want to estimate the values of the spikes \( \tilde{a}_i, 1 \leq i \leq r \) and \( \tilde{b}_j, 1 \leq j \leq s \), given the data matrix \( Y \) and the matrices \( A \) and \( B \). By Theorem 2.20 we see that \( \tilde{a}_i \) and \( \tilde{b}_j \) can be well approximated by \( \Omega_B(\tilde{\lambda}_{\pi_a(i)}) \) and \( \Omega_A(\tilde{\lambda}_{\pi_a(i)}) \), respectively. Together with Remark 2.21 we can propose the following estimators:
\[
\tilde{a}_i = \text{tr} A - \frac{1}{N} \sum_{i=r+s+1}^{N} \frac{a_i}{\lambda_{\pi_a(i)} G_{ii}(\lambda_{\pi_a(i)})} + 1,
\]
\[
\tilde{b}_j = \text{tr} B - \frac{1}{N} \sum_{j=r+s+1}^{N} \frac{b_j}{\lambda_{\pi_b(j)} G_{jj}(\lambda_{\pi_b(j)})} + 1.
\]
for \( 1 \leq i \leq r, 1 \leq j \leq s \). Note that our proposed estimators only need the information of \( A \) and \( B \) once we obtain the data matrix \( Y \). It is easy to see the above statistics are consistent estimators for the spikes.
Our results can also be used to estimate the number of spikes in \( r \), i.e., the values of \( r \) and \( s \). The number of spikes have important meanings in practice, for instance, it represents the number of factors in factor model and the number of signals in signal processing. In our model, we have two resources of spikes from either \( A \) or \( B \). We can follow the arguments of Section 4.1 of [20] to propose the estimators. For simplicity, we assume that Assumption 2.22 holds with \( \tau_1 = 1/3 \) and \( \tau_2 = 1/2 \). In this sense, by Theorem 2.23 and Remark 2.24 we see that

\[
|\hat{u}_k|^2 = \|k - \pi_a(i)\|_2 \frac{(\Omega_B^{-1})'(\hat{a}_i)}{\Omega_B^{-1} \hat{a}_i} + o(1).
\]

Similar discussion applies to \( \hat{v}_k \)'s. Therefore, for some constant \( c > 0 \), we propose the following statistics to estimate \( r \) and \( s \), respectively

\[
\hat{r} = \arg \min_{1 \leq i \leq cN} \left\{ \max_k |\hat{u}_i(k)|^2 \leq \omega \right\}, \quad \hat{s} = \arg \min_{1 \leq i \leq cN} \left\{ \max_k |\hat{v}_i(k)|^2 \leq \omega \right\},
\]

where the threshold \( \omega = o(1) \) can be chosen using a resampling procedure as discussed in [20] Section 4.1. Similar to the discussion of Theorem 4.3 of [20], we can conclude that \( \hat{r} \) and \( \hat{s} \) are consistent estimators for \( r \) and \( s \), respectively.

Finally, we mention that our results can be used for other statistical applications. For instance, the truncated Haar random matrices, i.e., submatrices of Haar random matrices, are important objects in the study of random sketching [22]. For such a generalization, we need to allow \( A \) or \( B \) be nonnegative in the sense that a portion of the eigenvalues of \( A \) or \( B \) are zeros for our model (1.1). In fact, all the results obtain in this paper still hold true with minor modifications. These results can deepen the understanding and provide more insights for the problem considered in [22], where only the global laws and convergent limits were used as the technical inputs. Moreover, in [15], the universality has been established for the local spectral statistics of the additive model in the bulk. For Wigner matrices and Gram matrices, the edge universality have been established in [32, 21]. The key technical inputs are the local laws. Therefore, we believe that combining the arguments in the aforementioned works and Theorem 2.12 we should be able to derive the distributions of the edge eigenvalues and use it for hypothesis testing regarding the models (1.1) and (1.3). We will pursue these directions in the future works.

### 3 Properties of subordination functions

In this section, we investigate the properties of the subordination functions. We first introduce some notations. The system of equations (2.8) can be written as

\[
\Phi_{\alpha\beta}(\Omega_\alpha(z), \Omega_\beta(z), z) = 0, \quad (\alpha, \beta) \in \mathbb{C}^2
\]

where we define \( \Phi_{\alpha\beta} \equiv (\Phi_\alpha, \Phi_\beta) : \{ (\omega_1, \omega_2, z) \in \mathbb{C}_+^3 : \arg \omega_1, \arg \omega_2 \geq \arg z \} \rightarrow \mathbb{C}^2 \) by

\[
\Phi_\alpha(\omega_1, \omega_2, z) := \frac{M_{\mu_\alpha} \omega_2 - \omega_1}{\omega_2 - z}, \quad \Phi_\beta(\omega_1, \omega_2, z) := \frac{M_{\mu_\beta} \omega_1}{\omega_2 - z}.
\]

Here \( \Phi_{\alpha\beta} \) is defined as a function of three complex variables. We will also use the following quantities, which are closely related to the first and the second derivatives of the system (3.1). Recall (2.4).

\[
S_{\alpha\beta}(z) := z^2 L'_{\mu_\beta}(\Omega_\alpha(z)) L'_{\mu_\alpha}(\Omega_\beta(z)) - 1,
\]

\[
T_\alpha(z) := \frac{1}{2} \left[ z L''_{\mu_\beta}(\Omega_\alpha(z)) L'_{\mu_\alpha}(\Omega_\beta(z)) + (z L'_{\mu_\beta}(\Omega_\alpha(z)))^2 L''_{\mu_\alpha}(\Omega_\beta(z)) \right],
\]

\[
T_\beta(z) := \frac{1}{2} \left[ z L''_{\mu_\alpha}(\Omega_\beta(z)) L'_{\mu_\beta}(\Omega_\alpha(z)) + (z L'_{\mu_\alpha}(\Omega_\beta(z)))^2 L''_{\mu_\beta}(\Omega_\alpha(z)) \right].
\]

13
We point out that, denoting by $D$ the differential operator with respect to $\omega_1$ and $\omega_2$, the first derivative of $\Phi_{\alpha \beta}$ is given by
\[
D\Phi_{\alpha \beta}(\omega_1, \omega_2, z) := \begin{pmatrix}
-\frac{1}{z} & L'_{\mu_\alpha}(\omega_2) \\
L'_{\mu_\beta}(\omega_1) & -\frac{1}{z}
\end{pmatrix},
\]
whose determinant is equal to $-z^{-2}S_{\alpha \beta}(z)$ at the point $(\Omega_\alpha(z), \Omega_\beta(z), z)$. Similarly, using $\Phi_{\alpha \beta}(\Omega_\alpha(z), \Omega_\beta(z), z) = 0$, we find that
\[
T_\alpha(z) = z \left[ \frac{\partial}{\partial \omega_1} \det D\Phi_{\alpha \beta}(\omega_1, z L_{\mu_\alpha}(\omega_1), z) \right]_{\omega_1 = \Omega_\alpha(z)},
\]
\[
T_\beta(z) = z \left[ \frac{\partial}{\partial \omega_2} \det D\Phi_{\alpha \beta}(z L_{\mu_\alpha}(\omega_2), \omega_2, z) \right]_{\omega_2 = \Omega_\beta(z)}.
\]
By replacing the pair $(\alpha, \beta)$ with $(A, B)$, we can define $\Phi_{AB}, S_{AB}, T_A, \text{ and } T_B$ analogously. The main result of this section is the following proposition, which establishes the properties of $\mu_A \boxtimes \mu_B$ and its associated subordination functions. For $z = E + i \eta \in \mathbb{C}_+$, we define
\[
\kappa \equiv \kappa(z) := |E - E_+|.
\]

**Proposition 3.1.** Suppose Assumptions 2.3 and 2.4 hold. Then for any fixed small constant $\tau > 0$ and sufficiently large $N$, the following hold:

(i) There exists some constant $C > 1$ such that
\[
\min_i |a_i - \Omega_\beta(z)| \geq C^{-1}, \quad \min_i |b_i - \Omega_\alpha(z)| \geq C^{-1},
\]
\[
C^{-1} \leq |\Omega_\alpha(z)| \leq C, \quad C^{-1} \leq |\Omega_\beta(z)| \leq C,
\]
uniformly in $z \in D_\tau(\eta_L, \eta_U)$.

(ii) For all $z \in D_\tau(\eta_L, \eta_U)$ and $\kappa$ defined in (3.6), we have
\[
\text{Im} m_{\mu_A \boxtimes \mu_B}(z) \sim \begin{cases}
\sqrt{\kappa + \eta}, & \text{if } E \in \text{supp } \mu_A \boxtimes \mu_B, \\
\frac{\eta}{\sqrt{\kappa + \eta}}, & \text{if } E \notin \text{supp } \mu_A \boxtimes \mu_B.
\end{cases}
\]

(iii) For all $z \in D_\tau(\eta_L, \eta_U)$, we have the following bounds for $S_{AB}, T_A, \text{ and } T_B$,
\[
S_{AB} \sim \sqrt{\kappa + \eta}, \quad |T_A(z)| \leq C, \quad |T_B(z)| \leq C.
\]
Furthermore, if $|z - E_+| \leq \delta$ for sufficiently small constant $\delta > 0$, we also have the lower bounds for $T_A$ and $T_B$:
\[
|T_A(z)| \geq c, \quad |T_B(z)| \geq c,
\]
where $c > 0$ is some constant.

(iv) For the derivatives of $\Omega_\alpha, \Omega_\beta$ and $S_{AB}$, we have
\[
|\Omega'_\alpha(z)| \leq C \frac{1}{\sqrt{\kappa + \eta}}, \quad |\Omega'_\beta(z)| \leq C \frac{1}{\sqrt{\kappa + \eta}}, \quad |S'_{AB}(z)| \leq C \frac{1}{\sqrt{\kappa + \eta}},
\]
uniformly in $z \in D_\tau(\eta_L, \eta_U)$.

The proof is divided into two parts. In Section 3.1, we prove an analogue of Proposition 3.1 for $\Omega_\alpha, \Omega_\beta$ and $\mu_\alpha \boxtimes \mu_\beta$ in Lemma 3.2 and Proposition 3.3, while in Section 3.2, by establishing the bounds $|\Omega_\alpha(z) - \Omega_\alpha(z)| = O_{\prec}(N^{-1/2})$ and $|\Omega_\beta(z) - \Omega_\beta(z)| = O_{\prec}(N^{-1/2})$, we extend the results of Section 3.1 to $\Omega_\alpha$ and $\Omega_\beta$, and complete the proof of Proposition 3.1. In Section 3.3, we prove the controls for the subordination functions, namely the upper bounds of $|\Omega_\alpha(z) - \Omega_\alpha(z)|$ and $|\Omega_\beta(z) - \Omega_\beta(z)|$. Proof therein follow the idea of [3], in the sense that we first prove the bound when $\text{Im} z$ is large and then use a bootstrapping argument to expand the domain of $z$. 
3.1 Free convolution of $\mu_\alpha \boxtimes \mu_\beta$

In this section, we collect the results concerning the $N$-independent measure $\mu_\alpha \boxtimes \mu_\beta$ and its corresponding subordination functions. Most of the results have been proved in [27] Section 5.

Lemma 3.2 (Lemma 5.2 and Proposition 5.6 of [27]). Suppose that $\mu_\alpha$ and $\mu_\beta$ satisfy Assumption 2.2. Then the following statements hold:

(i) For any compact $D \subset \mathbb{C}_+ \cup (0, \infty)$, there exists some constant $C > 0$ such that for all $z \in D$,
\[
C^{-1} \leq |\Omega_\alpha(z)| \leq C,
\]
\[
C^{-1} \leq |\Omega_\beta(z)| \leq C.
\]

(ii). There exists some constant $\varsigma > 0$ such that for all $z \in \mathbb{C}_+$,
\[
\text{dist}(\Omega_\alpha(z), \text{supp } \mu_\beta) \geq \varsigma,
\]
\[
\text{dist}(\Omega_\beta(z), \text{supp } \mu_\alpha) \geq \varsigma.
\]

(iii). Furthermore, we have
\[
0 < \Omega_\alpha(E_-) < E_\beta < E_\alpha < \Omega_\alpha(E_+),
\]
\[
0 < \Omega_\beta(E_-) < E_\alpha < E_\beta < \Omega_\beta(E_+).
\]

The lower bound in the second statement of Lemma 3.2 is the so-called stability bound, as it implies the stability of the systems (2.8) and (3.1). In particular, the edges of $\mu_\alpha \boxtimes \mu_\beta$ are completely characterized by the following equation. Recall the definitions of $E_-$ and $E_+$ of $\rho$ defined in 2.12.

Lemma 3.3 (Proposition 5.7 of [27]). The edges $E_-$ and $E_+$ satisfy the following equation when $z = E_\pm$
\[
\left( \frac{\Omega_\beta(z)}{M_{\mu_\alpha}(\Omega_\alpha(z))} M'_{\mu_\alpha}(\Omega_\alpha(z)) - 1 \right) \left( \frac{\Omega_\alpha(z)}{M_{\mu_\beta}(\Omega_\beta(z))} M'_{\mu_\beta}(\Omega_\beta(z)) - 1 \right) - 1 = 0.
\]

Remark 3.4. In fact the left-hand side of (3.9) is exactly $S_{\alpha\beta}(z)$. To see this, by the definitions of $L_{\mu_\alpha}(z)$ and $L_{\mu_\beta}(z)$ together with (2.8), we have
\[
L'_{\mu_\beta}(\Omega_\alpha(z)) = \frac{\Omega_\alpha(z) M'_{\mu_\beta}(\Omega_\alpha(z)) - M_{\mu_\beta}(\Omega_\alpha(z))}{\Omega_\alpha(z)^2} = \frac{\Omega_\alpha(z)}{z \Omega_\alpha(z)} \left[ \frac{\Omega_\alpha(z)}{M_{\mu_\beta}(\Omega_\beta(z))} M'_{\mu_\beta}(\Omega_\beta(z)) - 1 \right].
\]

Similar results hold for $L'_{\mu_\alpha}(\Omega_\beta(z))$ by interchanging $\alpha$ and $\beta$. In this sense, the edges of $\rho$ satisfy that $S_{\alpha\beta}(E_\pm) = 0$.

In [27] Proposition 6.15, the author has proved that there are exactly two real positive solutions to the equation (3.3), denoted as $E_-$ and $E_+$, so that \{ $x \in \mathbb{R}_+ : \rho(x) > 0$ \} = $(E_-, E_+)$. Using this as an input, the author in [27] also shows that the subordination functions admit the following Pick representations. Analogous results hold for the $L$-transform.

Lemma 3.5 (Lemma 5.8 of [27]). Suppose that $\mu_\alpha$ and $\mu_\beta$ satisfy Assumption 2.2. Then there exist unique finite measures $\mu_\alpha$, $\mu_\beta$, $\mu_\alpha$, and $\mu_\beta$ on $\mathbb{R}_+$ such that the following hold:
\[
L_{\mu_\alpha}(z) = 1 + m_{\mu_\alpha}(z), \quad \mu_\alpha(\mathbb{R}_+) = \text{Var } [\mu_\alpha] = \int_{\mathbb{R}_+} x^2 d\mu_\alpha(x) - 1, \quad \text{supp } \mu_\alpha = \text{supp } \mu_\alpha.
\]
\[
\frac{\Omega_\alpha(z)}{z} = 1 + m_{\mu_\alpha}(z), \quad \mu_\alpha(\mathbb{R}_+) = \text{Var } [\mu_\alpha], \quad \text{supp } \mu_\alpha = \text{supp } \rho.
\]

Similar results hold true if we replace $\alpha$ with $\beta$.

Armed with Lemma 3.2 we can see that $S_{\alpha\beta}$ is locally quadratic for $z$ around the edges $E_+$ and $E_-$ and consequently, $\Omega_\alpha$ and $\Omega_\beta$ have square root behavior. Specifically, denote
\[
\tilde{z}_+(\Omega) := \Omega \frac{M_{\mu_\alpha}^{-1}(\mu_\beta)(\Omega)}{M_{\mu_\beta}(\Omega)}.
\]
From the results of [27] Section 6.3, we find that there exists some neighborhood $U$ around $E_+$ such that when $z \in U$, $\tilde{z}_+(\Omega_0(z)) = z$ and
\[
z - E_+ = \tilde{z}_+(\Omega_0(z)) - \tilde{z}_+(\Omega_0(E_+)) = \frac{1}{2} z''_+ (\Omega_0(E_+)) (\Omega_0(z) - \Omega_0(E_+))^2 + O(|\Omega_0(z) - \Omega_0(E_+)|^3). \tag{3.12}
\]
Moreover, we have that $z''_+ (\Omega_0(E_+)) = 0$ and $z''_+ (\Omega_0(E_+)) > 0$.

**Lemma 3.6.** Suppose Assumption 2.2 holds. Then there exist positive constants $\gamma^\alpha_+, \gamma^\beta_+, 0 < \tau < 1$ and $\eta_0$ such that
\[
\Omega_0(z) = \Omega_0(E_+) + \gamma^\alpha_+ \sqrt{z - E_+} + O(|z - E_+|^3/2), \tag{3.13}
\]
uniformly in $z \in \{ z \in \mathbb{C}_+ : E_+ - \tau \leq E \leq \tau^{-1}, \ 0 \leq \eta < \eta_0 \}$, where $\sqrt{-1} = i$. The same asymptotics holds with $\alpha$ replaced by $\beta$. Also, for any compact interval $[a, b] \subset (E_-, E_+)$, there exists a constant $c > 0$ such that
\[
\text{Im} \Omega_0(x) > c, \text{ Im} \Omega_0(x) > c, \tag{3.14}
\]
for all $x \in [a, b]$.

*Proof.* (3.13) has been proved in [27], Proposition 5.10. We prove (3.14) here. The lower bound in (3.13) is a consequence of the equality $\{ x \in \mathbb{R} : \rho(x) > 0 \} = (E_-, E_+)$. In fact, since the density is continuous, there must exist a constant $c > 0$ so that $\rho(x) > c$ holds for all $x \in [a, b]$. Since
\[
\text{Im} \Omega_0(x) \int_{\mathbb{R}_+} \frac{t}{|t - \Omega_0(x)|^2} d\mu_\beta(t) = \text{Im}(xm_\rho(x) + 1) = x\rho(x) > ca,
\]
we conclude that $\text{Im} \Omega_0(x)$ is bounded below for $x \in [a, b]$ using (i) and (ii) of Lemma 3.2.

Furthermore, the density of the $\mu_\alpha \otimes \mu_\beta$ also has the square root behavior.

**Lemma 3.7** (Theorem 3.3 of [27]). Suppose that (i) and (ii) of Assumption 2.2 hold. Then there exists a constant $C > 1$ such that for all $x \in [E_-, E_+]$,
\[
C^{-1} \leq \frac{\rho(x)}{\sqrt{(E_+ - x)(x - E_-)}} \leq C.
\]

To characterize the behavior of $S_{\alpha \beta}, T_\alpha$ and $T_\beta$, we need the following lemma.

**Lemma 3.8.** Suppose Assumption 2.2 holds. Then there exist positive constants $\tau$ and $\eta_0$ such that the following hold uniformly in $z \in D_+(0, \eta_0)$
\[
L_{\mu_\alpha}'(\Omega_\beta(z)) \sim M_{\mu_\alpha}'(\Omega_\beta(z)) \sim 1, \quad L_{\mu_\alpha}''(\Omega_\beta(z)) \sim M_{\mu_\alpha}''(\Omega_\beta(z)) \sim 1, \quad |\Omega_\alpha'(z)| \sim \frac{1}{|z - E_+|^{3/2}}, \quad |\Omega_\alpha''(z)| \sim \frac{1}{|z - E_+|^{3/2}}. \tag{3.15}
\]
Similar results hold true if we replace $\alpha$ with $\beta$. Without loss of generality, we set $\tau$ to be the same as defined in (3.14).

*Proof.* We start with the first term of (3.15). By differentiating the first term in (3.10), we find that
\[
L_{\mu_\alpha}'(z) = \int \frac{1}{(x - z)^2} d\tilde{\mu}_\alpha(x), \quad M_{\mu_\alpha}'(z) = 1 + \int \frac{x}{(x - z)^2} d\tilde{\mu}_\alpha(x). \tag{3.17}
\]
Then the proof follows from Remark 2.10 supp $\tilde{\mu}_\alpha = \text{supp} \mu_\alpha$ (c.f. (3.10)) and (3.7). The second term follows from a similar argument by differentiating (3.17) and we omit the details.
Recall that the following hold:}

\(\Omega'_{\alpha}(z) = \frac{1}{z'_{\alpha}(\Omega_{\alpha}(z))} \sim \frac{1}{|z - E_{\pm}|}, \quad \Omega''_{\alpha}(z) = -\frac{z''_{\alpha}(\Omega_{\alpha}(z))\Omega'_{\alpha}(z)^2}{z'_{\alpha}(\Omega_{\alpha}(z))} \sim \frac{1}{|z - E_{\pm}|^{3/2}}. \quad (3.19)\)

Finally, we investigate the properties of \(\Sigma_{\alpha \beta}, \mathcal{T}_{\alpha}, \) and \(\mathcal{T}_{\beta}\) around the edge (and far away from bulk). They are summarized in the following proposition, whose proof is given in Appendix D.

**Proposition 3.9.** Suppose that Assumption 2.3 holds. Then there exist constants \(\tau, \eta_0 > 0\) such that the following hold:

(i) We have

\[
\text{Im} M_{\mu_\alpha, \mu_\beta}(z) \sim \text{Im} m_{\mu_\alpha, \mu_\beta}(z) = \text{Im} \Omega_\alpha(z) \sim \text{Im} \Omega_\beta(z) \sim \begin{cases} \sqrt{\kappa + \eta} & \text{if } E \in [E_-, E_+], \\ \eta & \text{if } E \notin [E_-, E_+], \end{cases}
\]

uniformly in \(z = E + i \eta \in \mathcal{D}_\tau(0, \eta_0).\)

(ii) For all fixed \(\eta_U \geq \eta_U,\) there exists a constant \(C > 0\) such that

\[
|\Sigma_{\alpha \beta}(z) \sim \sqrt{\kappa + \eta}, \quad |\mathcal{T}_{\alpha}(z)| \leq C, \quad |\mathcal{T}_{\beta}(z)| \leq C \quad (3.20)
\]

hold uniformly in \(z \in \mathcal{D}_\tau(0, \eta_U).\)

(iii) There exists a constant \(\delta > 0\) such that

\[
\mathcal{T}_{\alpha}(z) \sim 1, \quad \mathcal{T}_{\beta}(z) \sim 1
\]

hold uniformly in \(z \in \{z \in \mathbb{C}_+ : |z - E_+| < \delta\}.\)

### 3.2 Free convolution of \(\mu_A \boxtimes \mu_B: \text{proof of Proposition 3.1} \)

In this section we prove Proposition 3.1. In fact, we will prove our results on a slightly larger domain \(\mathcal{D}_0,\) where \(\mathcal{D}_0 := \mathcal{D}_I \cup \mathcal{D}_O\) and

\[
\mathcal{D}_I := \{z \in \mathbb{C}_+ : \text{Re } z \in [E_+ - \tau, E_+ + N^{-1+\xi \epsilon}], \text{Im } z \in [N^{-1+\xi \epsilon}, \eta_1]\},
\]

\[
\mathcal{D}_O := \{z \in \mathbb{C}_+ : \text{Re } z \in [E_+ + N^{-1+\xi \epsilon}, \tau^{-1}], \text{Im } z \in (0, \eta_1]\}, \quad (3.21)
\]

where \(\xi > 1\) is a fixed constant, \(\eta_1 > \eta_U\) will be given later in the proof (see Lemma 3.13). Our proof relies on the control of the difference between \(\Omega_{\alpha_\beta}\) and \((\Omega_A, \Omega_B).\)

**Lemma 3.10.** Let \(\mu_\alpha, \mu_\beta, \mu_A, \) and \(\mu_B\) satisfy Assumptions 2.2 and 2.4. Then there exists some constant \(C > 0\) such that for sufficiently large \(N\) the following statements hold:

(i) For all \(z \in \mathcal{D}_0,\)

\[
|\Omega_A(z) - \Omega_\alpha(z)| + |\Omega_B(z) - \Omega_\beta(z)| \leq C \frac{N^{-1+\epsilon}}{\sqrt{|z - E_+|}} \leq N^{-1/2+\epsilon}. \quad (3.22)
\]
for all \( N \), we have

\[
|\text{Im} \Omega_A(z) - \text{Im} \Omega_\alpha(z)| + |\text{Im} \Omega_B(z) - \text{Im} \Omega_\beta(z)| \leq C \frac{N^{-1+\epsilon} \text{Im}(\Omega_\alpha(z) + \Omega_\beta(z)) + \text{Im} z}{\sqrt{|z - E_+|}}.
\] (3.23)

We first present the proof of Proposition 3.1 using Lemma 3.10 and postpone its proof to Section 3.3. Along the proof of Proposition 3.1, we require two additional lemmas. The first result is an analogue of (representation theorem) to establish the existence and uniqueness of

Proof.

The proof is similar to that of [27, Lemmas 6.2 and 6.14]. First, we use classical Nevanlinna-Pick representation theorem (see [27, Lemma 3.7]) to establish the existence and uniqueness of \( \tilde{\mu}_A \) and \( \tilde{\mu}_A \). Second, we use the fact that \( L_{\mu_1} \) and \( \Omega_A \) are analytic in the complements of \([a_N, a_1]\) and \([\inf \text{supp} \mu_A \otimes \mu_B, \sup \text{supp} \mu_A \otimes \mu_B] \) to get the inclusion of the supports. We omit further details here and refer to the proofs of [27, Lemmas 6.2 and 6.14].

We remark that by (3.24), we have

\[
\text{Im} \Omega_A(z) = \text{Im}(z + zm_{\mu_A}(z)) = \eta + \eta \int \frac{x}{|x - z|^2} d\tilde{\mu}_A(x) \geq \eta,
\]

(3.25)

\[
\text{Im} \frac{\Omega_A(z)}{z} = \frac{1}{|z|^2} \int d\tilde{\mu}_A \geq \text{Var} [\mu_A] \frac{\eta}{2(|z|^2 + \|A\|^2 \|B\|^2)},
\]

(3.26)

where in the second step of (3.26) we used (2.11).

We next introduce the second auxiliary lemma. We will frequently refer to this lemma whenever we need to bound the differences between different Stieltjes transforms in terms of \( \mathcal{L}(\mu_\alpha, \mu_A) + \mathcal{L}(\mu_\beta, \mu_B) \). Its proof can be found in Appendix D.

**Lemma 3.11.** Suppose that \( \mu_1 \) and \( \mu_2 \) satisfy Assumptions 2.2 and 2.4 Then there exist unique probability measures \( \tilde{\mu}_A \), \( \tilde{\mu}_B \), \( \tilde{\mu}_A \), and \( \tilde{\mu}_B \) on \( \mathbb{R}_+ \) such that the following hold:

\[
L_{\mu_1}(z) = 1 + m_{\tilde{\mu}_A}(z), \quad \tilde{\mu}_A(\mathbb{R}_+) = \text{Var} [\mu_1], \quad \text{supp} \tilde{\mu}_A \subset [a_N, a_1],
\]

\[
\frac{\Omega_A(z)}{z} = 1 + m_{\tilde{\mu}_B}(z), \quad \tilde{\mu}_B(\mathbb{R}_+) = \text{Var} [\mu_1], \quad \text{supp} \tilde{\mu}_B \subset [\inf \text{supp} \mu_A \otimes \mu_B, \sup \text{supp} \mu_A \otimes \mu_B].
\]

(3.24)

Similar results hold true if we replace \( a \) with \( b \).

**Proof.** The proof is similar to that of [27, Lemmas 6.2 and 6.14]. First, we use classical Nevanlinna-Pick representation theorem (see [27, Lemma 3.7]) to establish the existence and uniqueness of \( \tilde{\mu}_A \) and \( \tilde{\mu}_A \). Second, we use the fact that \( L_{\mu_1} \) and \( \Omega_A \) are analytic in the complements of \([a_N, a_1]\) and \([\inf \text{supp} \mu_A \otimes \mu_B, \sup \text{supp} \mu_A \otimes \mu_B] \) to get the inclusion of the supports. We omit further details here and refer to the proofs of [27, Lemmas 6.2 and 6.14].

We remark that by (3.24), we have

\[
\text{Im} \Omega_A(z) = \text{Im}(z + zm_{\mu_A}(z)) = \eta + \eta \int \frac{x}{|x - z|^2} d\tilde{\mu}_A(x) \geq \eta,
\]

(3.25)

\[
\text{Im} \frac{\Omega_A(z)}{z} = \frac{1}{|z|^2} \int d\tilde{\mu}_A \geq \text{Var} [\mu_A] \frac{\eta}{2(|z|^2 + \|A\|^2 \|B\|^2)},
\]

(3.26)

where in the second step of (3.26) we used (2.11).

We next introduce the second auxiliary lemma. We will frequently refer to this lemma whenever we need to bound the differences between different Stieltjes transforms in terms of \( \mathcal{L}(\mu_\alpha, \mu_A) + \mathcal{L}(\mu_\beta, \mu_B) \). Its proof can be found in Appendix D.

**Lemma 3.12.** Let \( \delta > 0 \) be fixed and \( f \rightarrow \mathbb{C} \) be a continuous function which is continuously differentiable in \((E^\alpha - \delta, E^\alpha + \delta)\). Suppose that there exists \( N_0 \in \mathbb{N} \) independent of \( f \) such that \( d \leq \delta/2 \) and \([a_N, a_1] \in [E^\beta - \delta/2, E^\beta + \delta/2] \) for all \( N \geq N_0 \). Then there exists some constant \( C > 0 \), independent of \( f \), such that we have

\[
\left| \int_{\mathbb{R}_+} f(x) d\mu_\alpha(x) - \int_{\mathbb{R}_+} f(x) d\mu_\beta(x) \right| \leq C \|f'\|_{\text{Lip}, \delta} d,
\]

for all \( N \geq N_0 \), where we denoted

\[
\|f'\|_{\text{Lip}, \delta} := \sup_{x \in [E^\alpha - \delta, E^\alpha + \delta]} |f'(x)| + \sup \left\{ \left| \frac{f'(x) - f'(y)}{x - y} \right| : x, y \in [E^\alpha - \delta, E^\alpha + \delta], x \neq y \right\}.
\]

The same result holds if we replace \( \alpha \) and \( A \) by \( \beta \) and \( B \).

Now we are now ready to present the proof of Proposition 3.1.
Proof of Proposition 3.1. Throughout the proof we choose \( \eta_l \geq \eta_U \) such that \( D_\tau(\eta_L, \eta_U) \subset D_0 \). It is easy to see that (i) follows directly from Lemmas 3.2 and 3.10.

Next, we proceed to prove (ii). By (3.23), it is easy to verify that
\[
\text{Im} \, z \mu A \alpha B (z) = \text{Im} \, \Omega_B(z) \mu A (\Omega_B(z)) = \text{Im} \, \Omega_B(z) \int \frac{x}{|x - \Omega_B(z)|^2} d\mu_A(x).
\]
Together with (i) of Proposition 3.1 we obtain that
\[
\text{Im} \, z \mu A \alpha B (z) \sim \text{Im} \, \Omega_B(z). \tag{3.27}
\]

On the other hand, by (3.23), it is easy to verify that
\[
|\Omega_\beta (z) - \Omega_B(z)| \leq C N^{-1+\epsilon} \frac{1}{\sqrt{\kappa + \eta}} \leq \begin{cases} \frac{\sqrt{\kappa + \eta}}{\sqrt{\kappa + \eta}} & \text{if } E \in [E_-, E_+], \\ \frac{\eta}{\sqrt{\kappa + \eta}} & \text{if } E > E_+ \end{cases}
\]
holds uniformly in \( z \in D_0 \) when \( N \) is sufficiently large. In light of (3.27) and (i) of Proposition 3.9 we find that
\[
\text{Im} \, z \mu A \alpha B (z) \sim \text{Im} \, \Omega_\beta(z). \tag{3.29}
\]

On the other hand, by (2.10) and (i) of Proposition 3.1
\[
\text{Im} (z \mu A \alpha B (z)) = \eta \int \frac{x}{|x - z|^2} d(\mu_A \otimes \mu B)(x) \sim \eta \int \frac{1}{|x - z|^2} d(\mu_A \otimes \mu B)(x) = \text{Im} \, z \mu A \alpha B (z).
\]
Together with (3.29) and (i) of Proposition 3.9 we finish the proof of (ii).

For the proof of (iii), due to similarity, we will focus our proof on \( S_{AB}(z) \) and briefly discuss the proof of \( T_A(z) \) and \( T_B(z) \). Notice that
\[
S_{AB} - S_{\alpha \beta} = z^2 \left[ \frac{1}{|\Omega_\beta - \Omega_B(z)|^2} \left( L'_{\mu A} (\Omega_\alpha(z)) - L'_{\mu B} (\Omega_\alpha(z)) \right) \right. \left. + L'_{\mu B} (\Omega_\alpha(z)) \right. \left. \left( L'_{\mu A} (\Omega_\beta(z)) - L'_{\mu A} (\Omega_B(z)) \right) \right]. \tag{3.30}
\]
Hence, we need to control the right-hand side of (3.30). On one hand, we find that
\[
|L'_{\mu B} (\Omega_\alpha(z)) - L'_{\mu B} (\Omega_\alpha(z))| \leq |L'_{\mu A} (\Omega_\alpha(z)) - L'_{\mu A} (\Omega_\alpha(z))| + |L'_{\mu B} (\Omega_\alpha(z)) - L'_{\mu B} (\Omega_\alpha(z))|. \tag{3.31}
\]
By definition, we have that
\[
|L'_{\mu B} (\Omega_\alpha(z)) - L'_{\mu B} (\Omega_\alpha(z))| = \frac{1}{|\Omega_\alpha(z)^2 (\Omega_\alpha(z)) m_{\mu B} (\Omega_\alpha(z)) + 1)^{-2} \int \frac{x^2}{|x - \Omega_\alpha(z)|^2} d\mu_B (z) \tag{3.32}
\]
\[
- (\Omega_\alpha(z) m_{\mu B} (\Omega_\alpha(z)) + 1)^{-2} \int \frac{x^2}{|x - \Omega_\alpha(z)|^2} d\mu_B (x).
\]
By using Lemma 3.12 with \( f = (x - \Omega_\alpha(z))^2 \) and (iv) of Assumption 2.4 when \( N \) is large enough, we have
\[
L'_{\mu B} (\Omega_\alpha(z)) = L'_{\mu A} (\Omega_\alpha(z)) + O(N^{-1+\epsilon}),
\]
where we used (i) of Lemma 3.2 and (i) of Proposition 3.1. Together with (3.15), we can see that \( L''_{\mu B} (\omega) \sim 1 \) when \( \omega \) is around \( \Omega_\alpha(z) \) and \( N \) is large enough. Then by (3.32), it is easy to see that for some constant \( C > 0 \), we have
\[
|L'_{\mu B} (\Omega_\alpha(z)) - L'_{\mu B} (\Omega_\alpha(z))| \leq C |\Omega_\alpha(z) - \Omega_A(z)|.
\]
As a consequence, by (3.22), we can bound (3.31) by
\[
|L'_{\mu B} (\Omega_\alpha(z)) - L'_{\mu B} (\Omega_\alpha(z))| \leq C (|\Omega_\alpha(z) - \Omega_A(z)| + d) \leq C \frac{d}{\sqrt{\kappa + \eta}} \leq \sqrt{\kappa + \eta} \sim S_{\alpha \beta}(z). \tag{3.33}
\]
where we used the definition of $D_\tau(\eta_L, \eta_U)$ and choose $\epsilon < \gamma$ in the third inequality. Note that $L_{\mu_A}'(\Omega_B(z)) \sim 1$ by a discussion similar to (3.15). This finishes the proof of the first term of the right-hand side of (3.30).

Similar discussion can be applied to the second term. In summary, we have that

$$|S_{AB}(z) - S_{\alpha\beta}(z)| \ll S_{\alpha\beta}(z).$$

(3.34)

The we complete the proof of the control of $S_{AB}(z)$ in (iii) using (ii) of Proposition 3.9. For the control of $T_A$ and $T_B$, by the definition of $T_\alpha$ in (3.4), with an argument similar to (3.34), we find that $T_A - T_\alpha = o(1)$ and $T_B - T_\beta = o(1)$. This concludes the proof using (ii) and (iii) of Proposition 3.9.

Finally, we prove (iv). By applying $\frac{d}{dz}$ to the subordination system (3.1) with $\alpha$ and $\beta$ replaced by $A$ and $B$, we have

$$L_{\mu_A}'(\Omega_B(z))\Omega_B'(z) - \frac{\Omega_A'(z)}{z} + \frac{\Omega_A(z)}{z^2} = 0,$$

$$L_{\mu_B}'(\Omega_A(z))\Omega_A'(z) - \frac{\Omega_B'(z)}{z} + \frac{\Omega_B(z)}{z^2} = 0.$$

Equivalently, we rewrite the above equations as

$$\begin{pmatrix} zL_{\mu_B}'(\Omega_A(z)) & -1 \\ -1 & zL_{\mu_A}'(\Omega_B(z)) \end{pmatrix} \begin{pmatrix} \Omega_A'(z) \\ \Omega_B'(z) \end{pmatrix} = -\frac{1}{z} \begin{pmatrix} \Omega_A(z) \\ \Omega_B(z) \end{pmatrix}.$$

Note that the determinant of $(2 \times 2)$-matrix on the left-hand side of the above equation is $S_{AB}(z)$. Then we have

$$\begin{pmatrix} \Omega_A'(z) \\ \Omega_B'(z) \end{pmatrix} = -\frac{1}{zS_{AB}(z)} \begin{pmatrix} zL_{\mu_B}'(\Omega_B(z)) & 1 \\ 1 & zL_{\mu_A}'(\Omega_A(z)) \end{pmatrix} \begin{pmatrix} \Omega_B(z) \\ \Omega_A(z) \end{pmatrix}.$$

(3.35)

By a discussion similar to (3.15), (i) of Lemma 3.2 and (i) of Proposition 3.1, we find that the entries of in the brackets of the right-hand side of (3.35) are bounded from above and below. Then by (iii) of Proposition 3.1, we get the desired result of $\Omega_A'(z)$ and $\Omega_B'(z)$. For $S_{AB}'(z)$, we first note that

$$S_{AB}'(z) = \frac{2}{z}(S_{AB} + 1) + z^2L_{\mu_B}''(\Omega_A(z))L_{\mu_A}'(\Omega_B(z))\Omega_A'(z) + z^2L_{\mu_B}'(\Omega_A(z))L_{\mu_A}''(\Omega_B(z))\Omega_B'(z).$$

(3.36)

Since

$$\frac{2}{z}(S_{AB} + 1) = 2zL_{\mu_A}'(\Omega_B(z))L_{\mu_B}'(\Omega_A(z)),$$

by a discussion similar to (3.15), we find that

$$\frac{2}{z}(S_{AB} + 1) \sim 1 \leq C\frac{1}{\sqrt{\kappa + \eta}}, z \in D_\tau(\eta_L, \eta_U).$$

Similarly, we can bound the other terms on the right-hand side of (3.36) using the first two terms of (iv) of Proposition 3.1. This finishes the proof of the bound for $S_{AB}'(z)$.

$$\square$$

3.3 Closeness of subordination functions: proof of Lemma 3.10

In this section, we prove Lemma 3.10. By the construction of (3.1), we find that $\Omega_A(z)$ and $\Omega_B(z)$ are determined by the equation

$$\Phi_{AB}(\Omega_A, \Omega_B, z) = 0.$$

Equivalently, $\Omega_A$ and $\Omega_B$ are solutions of the equations

$$\Phi_{\alpha}(\omega_1, \omega_2, z) = r_1(z), \quad \Phi_{\beta}(\omega_1, \omega_2, z) = r_2(z),$$

(3.37)
where $\Phi_\alpha$ and $\Phi_\beta$ are defined in (3.2) and
\[
  r_1(z) := \frac{M_{\mu_\alpha}(\omega_2) - M_{\mu_A}(\omega_2)}{\omega_2}, \quad r_2(z) := \frac{M_{\mu_\beta}(\omega_1) - M_{\mu_B}(\omega_1)}{\omega_1}. \tag{3.38}
\]
In what follows, we use the notations $r_A(z) \equiv r_1(\Omega_B(z))$ and $r_B(z) \equiv r_2(\Omega_A(z))$. Ideally, both $r_A(z)$ and $r_B(z)$ should be small and we aim to bound $|\Omega_A - \Omega_\alpha(z)|$ and $|\Omega_B(z) - \Omega_\beta(z)|$ in terms of the norm of $r(z) := (r_A(z), r_B(z))$.

Since the lower bounds of $\text{dist}(\Omega_A, \mu_\beta)$ and $\text{dist}(\Omega_B, \mu_\alpha)$ play an important role in our proof, in light of (3.25), we split the proof of Lemma 3.10 into two steps. First, we start from the regime in which $\eta = \text{Im} z$ is large. Second, we use a bootstrapping argument to extend the result to whole domain $D_0$. We mention that when $\eta$ is large, many critical quantities can be bounded from above and below. For instance, when $\eta = O(1)$, we assert that $\Omega_A(z)$ and $\Omega_B(z)$ are bounded from below in light of (3.26).

In the first step, we will need the following lemma. Its proof relies on the a consequence of Kantorovich theorem (c.f. Lemma 3.14), which provides an upper bound for deviation between the initial value and the exact solution is an application of Newton’s method. In our case, we consider $(\Omega_A(z), \Omega_B(z))$ as the initial point to find the solution $(\Omega_\alpha(z), \Omega_\beta(z))$ of $\Phi_{\alpha\beta}(\cdot, \cdot, z) = 0$. The upper bound in Kantorovich’s theorem is given in terms of the first and second derivatives of $\Phi_{\alpha\beta}(\cdot, \cdot, z)$. We further show that they can be bounded by $2 \|r(z)\|$. It will be proved in Appendix D.

**Lemma 3.13.** There exists a constant $\eta_1 > 0$ such that for all $z \in \mathbb{C}_+$ with $\text{Re} z \in [E_+ - \tau, \tau^{-1}]$ and $\text{Im} z = \eta_1$ the following hold:
\[
  |\Omega_\alpha(z) - \Omega_A(z)| \leq 2 \|r(z)\|, \quad |\Omega_\beta(z) - \Omega_B(z)| \leq 2 \|r(z)\|,
\]
for all sufficiently large $N$.

For the second step, we will need the following lemma to extend the results in Lemma 3.13 for smaller $\text{Im} z$. Denote $K_1 := 4K_4$ and $K_2 := (4K_3K_4)^{-1}$, where
\[
  K_3 := 27\kappa_0^{-3} \max(\hat{\mu}_\alpha(\mathbb{R}_+), \hat{\mu}_\beta(\mathbb{R}_+)),
\]
with $\kappa_0$ defined as
\[
  \kappa_0 = \min_{z \in \mathbb{C}_+} \{\text{dist}(\Omega_\alpha(z), \text{supp} \mu_\beta), \text{dist}(\Omega_\beta(z), \text{supp} \mu_\alpha)\},
\]
and
\[
  K_4 := \sup \left\{ |z| + |z^2 L'_{\mu_\alpha}(\Omega_\beta(z))|, |z| + |z^2 L'_{\mu_\beta}(\Omega_\alpha(z))| : z \in D_\tau(0, \eta_1) \right\},
\]
where $\eta_1$ is introduced in Lemma 3.13. It is easy to see that both $K_3$ and $K_4$ are positive finite numbers. We next state the results. Recall (3.14).

**Lemma 3.14.** For sufficiently large $N$ and $z_0 \in D_\tau(0, \eta_1)$, where $\eta_1$ is introduced in Lemma 3.14, suppose that there exists some $q > 0$ such that the following hold:

(i). $q \leq \frac{1}{3} \kappa_0$;

(ii). $|\Omega_A(z_0) - \Omega_\alpha(z_0)| \leq q, \quad |\Omega_B(z_0) - \Omega_\beta(z_0)| \leq q$;

(iii). $q \leq \frac{1}{2} K_2 S_{\alpha\beta}(z_0)$.

Then we have
\[
  |\Omega_A(z_0) - \Omega_\alpha(z_0)| + |\Omega_B(z_0) - \Omega_\beta(z_0)| \leq K_1 \frac{\|r(z_0)\|}{S_{\alpha\beta}(z_0)}.
\]
The proof of Lemma 3.14 will be provided in Appendix D. Given these two lemmas, we proceed to finish the proof of Lemma 3.10 using the two-step strategy as described earlier.

**Proof of Lemma 3.10.** We start with the proof of (3.22). We first prove that there exists a constant $C_0 > 0$ and $N_0 \in \mathbb{N}$ such that the following holds; if $N \geq N_0$ and $z \in \mathcal{D}_0$ satisfy the assumptions of Lemma 3.14, then

$$K_1 \|r(z)\| \leq C_0 \frac{N^{-1+\epsilon/2}}{\sqrt{K + \eta}}.$$

Note that (3.7) ensures that at least one of the following inequalities should hold for any $z \in \mathcal{D}_0$:

$$\text{Re} \Omega(z) \leq E_0^\alpha - \frac{\kappa_0}{2}, \quad \text{Re} \Omega(z) \geq E_0^\alpha + \frac{\kappa_0}{2}, \quad \text{Im} \Omega(z) \geq \frac{\kappa_0}{2}.$$  \hspace{1cm} (3.39)

Moreover, by assumptions (i) and (ii) of Lemma 3.14, $\Omega(z)$ satisfies

$$\text{Re} \Omega(z) \leq E_0^\alpha - \frac{\kappa_0}{6}, \quad \text{Re} \Omega(z) \geq E_0^\alpha + \frac{\kappa_0}{6}, \quad \text{Im} \Omega(z) \geq \frac{\kappa_0}{6}.$$  \hspace{1cm} (3.39)

Thus, by Lemma 3.12 we have

$$\left| \int \frac{xd(\mu_A - \mu_B)(x)}{x - \Omega_B(z)} \right| \leq C d,$$  \hspace{1cm} (3.40)

where $C$ is some constant. Furthermore, we have

$$\left| \int \frac{x}{x - \Omega_B(z)} d\mu_A(x) \right| \geq \left| \int \frac{x}{x - \Omega_B(z)} d\mu_A(x) \right| - \int \frac{x}{x - \Omega_B(z)} d\mu_B(x) \geq 0.$$  \hspace{1cm} (3.41)

for some positive constant $c > 0$. Together with (D.5), we find that $|r_A(z)| \leq C d$. Similarly, we can show $|r_B(z)| \leq C d$ and consequently, $\|r(z)\| \leq C d$. This implies

$$K_1 \|r(z)\| \leq C_0 \frac{N^{-1+\epsilon/2}}{\sqrt{K + \eta}}.$$  \hspace{1cm} (3.42)

Now we prove (3.22) using the result above. First, we fix $N \geq N_0$ and $z \in \mathcal{D}_0$. Define a finite decreasing sequence $(\eta_1, \cdots, \eta_M)$ of positive numbers, starting from $\eta_1$ given in Lemma 3.13 and ending with $\eta_M = \eta$, such that

$$\eta_i - \eta_{i+1} \leq L := \eta^2 N^{-(1+1/\epsilon)/2}.$$  \hspace{1cm} (3.43)

Note that the length $M$ of this sequence may depend on $N$. Then it suffices to prove that $z_i = E + i\eta_i$ satisfies the assumption of Lemma 3.14 for each $i = 1, \cdots, M$. Suppose inductively that the result holds for some $i$. By Lemmas 3.5 and 3.11 we have

$$|\Omega_A(z_i) - \Omega_A(z_{i+1})| \leq C_1 \eta_i^{-2} L,$$

for some constant $C_1 > 0$ that depends only on $\mu_A$, and the same bound holds also for $\Omega_B$ and $\Omega_B$. Now we see that

$$|\Omega_A(z_{i+1}) - \Omega_A(z_{i}) + \Omega_B(z_{i}) - \Omega_B(z_{i+1})| \leq |\Omega_A(z_i) - \Omega_A(z_{i+1})| + |\Omega_B(z_i) - \Omega_B(z_{i+1})| + 2C_1 \eta_i^{-2} L \leq K_1 \frac{\|r(z_i)\|}{\|S_{\alpha, \beta}(z_i)\|} + 2C_1 \eta_i^{-2} L \leq C_0 \frac{N^{-1+\epsilon/2}}{\sqrt{K + \eta_i}} + 2C_1 N^{-(1+1/\epsilon)/2} \leq (C_0 + 2C_1) N^{-(1+1/\epsilon)/2}.$$  \hspace{1cm} (3.44)

Thus, with the choice

$$q = (C_0 + 2C_1) N^{-(1+1/\epsilon)/2},$$

we can enlarge $N_0$ so that the following hold:

$$q = (C_0 + 2C_1) N^{-(1+1/\epsilon)/2} \leq \frac{1}{3},$$

for all $z \in \mathcal{D}_0$.
where we used $\xi > 1$. Therefore $z_{i+1}$ satisfies the assumptions of Lemma \ref{theo:3.12}. Since $z_1$ automatically satisfies the assumptions of Lemma \ref{theo:3.12} by Lemma \ref{theo:3.13}, we conclude the proof of (3.22) by induction.

We next prove (3.23). The proof is similar to that of Lemma \ref{theo:3.14} by considering the imaginary parts. The key ingredients are (3.22) and the fact
\[
\operatorname{Im} r_A(z) \leq C \operatorname{Im} \Omega_B(z) \left( d + \int \frac{x^2}{|x - \Omega_B(z)|^2} d(\mu_\alpha - \mu_A)(x) \right) \leq C \operatorname{Im} \Omega_B(z) d,
\]
where we use (3.22), (3.40), (3.41) and $\operatorname{Im} \frac{r}{z} = \operatorname{Im} z \frac{x^2}{|x - z|^2}$ and the definition $\operatorname{Im} r_A(z)$ is given by
\[
\operatorname{Im} r_A(z) = \operatorname{Im} \left( \int \frac{x \Omega_B(z)}{x - \Omega_B(z)} d\mu_\alpha(x) \right)^{-1} - \operatorname{Im} \left( \int \frac{x \Omega_B(z)}{x - \Omega_B(z)} d\mu_A(x) \right)^{-1}.
\]

Similar results hold for $\operatorname{Im} r_B(z)$. By taking imaginary parts of perturbed equation (3.37), we have
\[
\begin{align*}
\operatorname{Im} L_{\mu_\alpha}(\Omega_B(z)) - \operatorname{Im} \frac{\Omega_A(z)}{z} &= \operatorname{Im} r_A(z), \\
\operatorname{Im} L_{\mu_\beta}(\Omega_B(z)) - \operatorname{Im} \frac{\Omega_B(z)}{z} &= \operatorname{Im} r_B(z).
\end{align*}
\]

The rest of the proof follows from a discussion similar to the proof of Lemma \ref{theo:3.14}. We omit the details here. \hfill $\blacksquare$

### 4 Proof of Theorems 2.20 and 2.23

In this section, we prove the main results regarding the spiked invariant model in Section 2.3. For the convenience of our proof, we introduce the following linearization form. For $z \in \mathcal{D}_\tau(\eta_L, \eta_U) \cup \mathcal{D}_\tau(\eta_U)$ in (2.14) and (2.20). Recall $Y = A^{1/2}U B^{1/2}$. Denote $H = H(z)$ as
\[
H(z) := \begin{pmatrix}
0 & \frac{z^{1/2}Y}{\sqrt{2}} \\
\frac{z^{1/2}Y^*}{\sqrt{2}} & 0
\end{pmatrix},
\]
and $G(z) = (H - z)^{-1}$. By Schur’s complement, it is easy to see that
\[
G(z) = \begin{pmatrix}
\tilde{G}(z) & z^{-1/2}\tilde{G}(z)Y \\
z^{-1/2}\tilde{G}(z)^* & \tilde{G}(z)
\end{pmatrix},
\]
where we recall the definitions in (2.8). For simplicity of the notations, we define the index sets
\[
\mathcal{I}_1 := \{1, \ldots, N\}, \quad \mathcal{I}_2 := \{N + 1, \ldots, 2N\}, \quad \mathcal{I} := \mathcal{I}_1 \cup \mathcal{I}_2.
\]

Then we relabel the indices of the matrices according to
\[
U = (U_{i\mu} : i \in \mathcal{I}_1, \mu \in \mathcal{I}_2), \quad A = (A_{ij} : i, j \in \mathcal{I}_1), \quad B = (B_{\mu\nu} : \mu, \nu \in \mathcal{I}_2).
\]

In the proof of this section, we will consistently use the latin letters $i, j \in \mathcal{I}_1$ and greek letters $\mu, \nu \in \mathcal{I}_2$. We denote the $2N \times 2N$ diagonal matrix $\Theta \equiv \Theta(z)$ by letting
\[
\Theta_{ii} = \frac{1}{z} \frac{\Omega_B(z)}{A_i - \Omega_B(z)}, \quad \Theta_{\mu\nu} = \frac{1}{z} \frac{\Omega_A(z)}{B_{\mu\nu} - \Omega_A(z)}.
\]

Similar to Theorems 2.12 and 2.16, we have the following controls for the resolvent $G(z)$ uniformly in $z \in \mathcal{D}_\tau(\eta_L, \eta_U) \cup \mathcal{D}_\tau(\eta_U)$. It is the key technical input for this section. We postpone its proof to Section 8.3.
Proposition 4.1. Under the assumptions of Theorem 2.12, we have

\[
\sup_{1 \leq k, l \leq 2N} |(G(z) - \Theta(z))_{kl}| \prec \sqrt{\text{Im}\mu_A \S_{\mu_B} N + 1} N\eta
\]

holds uniformly in \(z \in \mathcal{D}_\tau(\eta_L, \eta_U)\). Moreover, when the assumptions of Theorem 2.16 hold, we have that

\[
\sup_{1 \leq k, l \leq 2N} |(G(z) - \Theta(z))_{kl}| \prec N^{-1/2} (\kappa + \eta)^{-1/4}, \quad \kappa = |z - E_+|,
\]

holds uniformly in \(z \in \mathcal{D}_\tau(\eta_U)\).

4.1 Eigenvalue statistics: proof of Theorem 2.20

We first provide the following lemma which gives the master equation for the locations of the outlier eigenvalues. Denote

\[
U = \begin{pmatrix} E_r & 0 \\ 0 & E_s \end{pmatrix}, \quad \mathcal{D} = \begin{pmatrix} D^a(D^a + 1)^{-1} & 0 \\ 0 & D^b(D^b + 1)^{-1} \end{pmatrix},
\]

where \(E_r = (e_1, \cdots, e_r), \quad E_s = (e_1, \cdots, e_s), \quad D^a = \text{diag}(d^a_1, \cdots, d^a_r)\) and \(D^b = \text{diag}(d^b_1, \cdots, d^b_s)\). For notational convenience, we set \(Q_1 = \tilde{H}\) as defined in (2.2).

Lemma 4.2. If \(x \neq 0\) is not an eigenvalue of \(Q_1\), then it is an eigenvalue of \(\tilde{Q}_1\) if and only if

\[
\det(\mathcal{D}^{-1} + z U^* G(x) U) = 0.
\]

Proof. See Lemma S.4.1 of [20].

Heuristically, by Proposition 4.1 and Lemma 4.2, an outlier location \(x > E_+\) should satisfy the condition that

\[
\prod_{i=1}^r \left( d^a_i + 1 + \frac{\Omega_B(x)}{a_i - \Omega_B(x)} \right) \prod_{j=1}^s \left( d^b_j + 1 + \frac{\Omega_A(x)}{b_j - \Omega_A(x)} \right) = 0.
\]

We write

\[
\frac{d^a_i + 1}{d^a_i} + \frac{\Omega_B(x)}{a_i - \Omega_B(x)} = \frac{1}{d^a_i} - \frac{a_i}{\Omega_B(x) - a_i}.
\]

Since \((\Omega_B(x) - a_i)^{-1}\) is decreasing (c.f. (3.19)), by (3.8), we find that when \(x > E_+\)

\[
\frac{d^a_i + 1}{d^a_i} + \frac{\Omega_B(x)}{a_i - \Omega_B(x)} = 0,
\]

if and only if

\[
\frac{d^a_i + 1}{d^a_i} + \frac{\Omega_B(E_+)}{a_i - \Omega_B(E_+)} < 0,
\]

which implies that \(\tilde{a}_i > \Omega_B(E_+)\). Similar calculation holds for \(\tilde{b}_j, 1 \leq j \leq s\). We now proceed to prove Theorem 2.20. We will follow the basic strategy as summarized in the beginning of [20, Appendix S.4] for the i.i.d. model, i.e., the entries of \(U\) are centered i.i.d. random variables with variance \(N^{-1}\). We emphasize that similar strategies have been employed in [13, Section 4] for spiked covariance matrices and in [20, Section 6] for finite rank deformation of Wigner matrices.
Proof of Theorem 2.20. By Proposition 4.4 and Theorem 2.14, for any fixed \( \epsilon > 0 \), we can choose a high probability event \( \Xi \equiv \Xi(\epsilon) \) where the following estimates hold:

\[
1(\Xi) \| U^*(G(z) - \Theta(z)) U \| \leq N^{\epsilon/2} \left( \frac{\text{Im} m_{\mu A} \delta_{\mu A}}{N \eta} + \frac{1}{N \eta} \right), \quad z \in \mathcal{D}_{\tau}(\eta_L, \eta_U); \tag{4.4}
\]

\[
1(\Xi) \| U^*(G(z) - \Theta(z)) U \| \leq N^{-1/2 + \epsilon/2}(\kappa + \eta)^{-1/4}, \quad z \in \mathcal{D}_{\tau}(\eta_U); \tag{4.5}
\]

\[
1(\Xi) |\lambda_i(Q_1) - E_+| \leq n^{-2/3 + \epsilon}, \quad 1 \leq i \leq \infty. \tag{4.6}
\]

We will restrict our proof to \( \Xi \) in what follows and hence the discussion below will be entirely deterministic. We first prepare some notations following the proof of [20, Theorem 3.6]. For any fixed constant \( \epsilon > 0 \), we denote

\[
\mathcal{O}_\epsilon^{(a)} := \left\{ i : \hat{a}_i - \Omega_B(E_+) \geq N^{-1/3+\epsilon} \right\}, \quad \mathcal{O}_\epsilon^{(b)} := \left\{ \mu : N + 1 \leq \mu \leq 2N, \hat{b}_\mu - \Omega_A(E_+) \geq N^{-1/3+\epsilon} \right\},
\]

Here and after, we use \( \hat{b}_\mu := \hat{b}_\mu - N \) for \( \mu \in I_2 \). Recall that the eigenvalues of \( \hat{A} \) and \( \hat{B} \) are ordered in the decreasing fashion. It is easy to see that

\[
\sup_{\mu \notin \mathcal{O}_\epsilon^{(b)}} (\hat{b}_\mu - \Omega_A(E_+)) \lesssim N^{-1/3+\epsilon}, \quad \inf_{\mu \in \mathcal{O}_\epsilon^{(b)}} (\hat{b}_\mu - \Omega_A(E_+)) \gtrsim N^{-1/3+\epsilon}.
\]

Moreover, since we are mainly interested in the outlying and extremal non-outlier eigenvalues, we use the convention that \( \Omega_B^{-1}(\hat{a}_i) = E_+, i \geq r \) and \( \Omega_A^{-1}(\hat{b}_\mu) = E_+, \mu \geq N + s \). Throughout the proof, we will need the following estimate following from [3,13] and [5,22]

\[
\Omega_B^{-1}(\hat{a}_i) - E_+ \sim (\hat{a}_i - \Omega_B(E_+))^2. \tag{4.7}
\]

Indeed, when \( \Omega_B^{-1}(\hat{a}_i) - E_+ \leq \zeta_1 \) for some sufficiently small constant \( 0 < \zeta_1 < 1 \), using [3,13], [5,22] and the fact \( \Omega_B(\cdot) \) is monotone increasing, we readily see that

\[
\hat{a}_i \sim \Omega_B(E_+) + \gamma \sqrt{\Omega_B^{-1}(\hat{a}_i) - E_+} + N^{-1/2+\epsilon}, \tag{4.8}
\]

where \( \gamma > 0 \) is some universal constant. This immediately implies [4,7] using Assumption 2.17. On the other hand, when \( \Omega_B^{-1}(\hat{a}_i) - E_+ \geq \zeta_1 \), since \( \Omega_B(\cdot) \) is increasing, we obtain that

\[
\hat{a}_i \geq \Omega_B(E_+ + \zeta_1) \sim \Omega_B(E_+) + \gamma' \sqrt{\zeta_1} + N^{-1/2+\epsilon}.
\]

This proves the claim [4,7].

As for any \( i \notin \mathcal{O}_\epsilon^{(a)} \) and \( \mu \in \mathcal{O}_\epsilon^{(b)} \), by (4.7), \( \Omega_B^{-1}(\hat{a}_i) \leq E_+ + N^{-2/3+2\epsilon} \leq \Omega_A^{-1}(\hat{b}_\mu) \), we have the followings

\[
\sup_{i \notin \mathcal{O}_\epsilon^{(a)}} \Omega_B^{-1}(\hat{a}_i) \leq \inf_{\mu \in \mathcal{O}_\epsilon^{(b)}} \Omega_A^{-1}(\hat{b}_\mu) + N^{-2/3+\epsilon}.
\]

Similarly, we have that

\[
\sup_{\mu \notin \mathcal{O}_\epsilon^{(b)}} \Omega_A^{-1}(\hat{b}_\mu) \leq \inf_{i \notin \mathcal{O}_\epsilon^{(a)}} \Omega_B^{-1}(\hat{a}_i) + N^{-2/3+\epsilon}.
\]

An advantage of the above labelling is that the largest outliers of \( \hat{Q}_1 \) can be labelled according to \( i \in \mathcal{O}_\epsilon^{(a)} \) and \( \mu \in \mathcal{O}_\epsilon^{(b)} \). Analogously to (S.9) and (S.10) of [20], we find that to prove Theorem 2.20 it suffices to prove that for arbitrarily small constant \( \epsilon > 0 \), there exists some constant \( C > 0 \) such that

\[
1(\Xi) \left| \hat{\lambda}_{m_\epsilon(i)} - \Omega_B^{-1}(\hat{a}_i) \right| \leq CN^{-1/2+2\epsilon} \Delta_1(\hat{a}_i), \quad 1(\Xi) \left| \hat{\lambda}_{m_\mu}(\mu) - \Omega_A^{-1}(\hat{b}_\mu) \right| \leq CN^{-1/2+2\epsilon} \Delta_2(\hat{b}_\mu), \tag{4.9}
\]
for all \( i \in \mathcal{O}^{(a)}_{4\epsilon} \) and \( \mu \in \mathcal{O}^{(b)}_e \), where we used the short-hand notations
\[
\Delta_1(\hat{a}_i) := (\hat{a}_i - \Omega_B(E_+))^{1/2}, \quad \Delta_2(\hat{b}_\mu) := (\hat{b}_\mu - \Omega_A(E_+))^{1/2},
\]
and we let \( \pi_b \) be defined on the set \( \{ N + 1, \ldots, 2N \} \) for notational convenience, and
\[
\mathbf{1}(\Xi) \left| \hat{\lambda}_{\pi_b(i)} - E_+ \right| \leq CN^{-2/3+12\epsilon}, \quad \mathbf{1}(\Xi) \left| \hat{\lambda}_{\pi_b(\mu)} - E_+ \right| \leq CN^{-2/3+12\epsilon},
\]
for all \( i \in \{1, 2, \ldots, r\} \setminus \mathcal{O}^{(a)}_{4\epsilon} \) and \( \mu \in \{ N + 1, \ldots, N + s \} \setminus \mathcal{O}^{(b)}_e \).

We now follow the strategy of [20, Theorem 3.6] to complete our proof. We divide our discussion into four steps for the convenience of the readers. We will focus on explaining Step 1 since it differs the most from its counterpart of the proof of [20, Theorem 3.6], and briefly sketch Steps 2-4.

**Step 1:** For each \( 1 \leq i \leq r^+ \), we define the permissible intervals
\[
I^{(a)}_i := \left[ \Omega_B^{-1}(\hat{a}_i) - N^{-1/2+\epsilon}\Delta_1(\hat{a}_i), \Omega_B^{-1}(\hat{a}_i) + N^{-1/2+\epsilon}\Delta_1(\hat{a}_i) \right].
\]
Similarly, for each \( 1 \leq \mu - N \leq s^+ \), we denote
\[
I^{(b)}_\mu := \left[ \Omega_A^{-1}(\hat{b}_\mu) - N^{-1/2+\epsilon}\Delta_2(\hat{b}_\mu), \Omega_A^{-1}(\hat{b}_\mu) + N^{-1/2+\epsilon}\Delta_2(\hat{b}_\mu) \right].
\]
Then we define
\[
I := I_0 \cup \left( \bigcup_{i \in \mathcal{O}^{(a)}_{4\epsilon}} I^{(a)}_i \right) \cup \left( \bigcup_{\mu \in \mathcal{O}^{(b)}_e} I^{(b)}_\mu \right), \quad I_0 := \left[ 0, E_+ + n^{-2/3+3\epsilon} \right].
\]

The main task of this step is to prove that on the event \( \Xi \), there exist no eigenvalues outside \( I \). This is summarized as the following lemma.

**Lemma 4.3.** The complement of \( I \) contains no eigenvalues of \( \hat{Q}_1 \).

**Proof.** By Lemma 4.2, 4.6, and 4.7, we find that \( x \notin I_0 \) is an eigenvalue of \( \hat{Q}_1 \) if and only if
\[
1(\Xi)(\mathbf{D}^{-1} + x U^* G(x) U) = 1(\Xi) \left( \mathbf{D}^{-1} + x U^* \Theta(x) U + O(\kappa^{-1/4} N^{-1/2+\epsilon/2}) \right),
\]
is singular. In light of 4.13, it suffices to show that if \( x \notin I \), then
\[
\min \left\{ \min_{1 \leq i \leq r} \left| \frac{d^2_i + 1}{d^2_i} + \frac{\Omega_B(x)}{a_i - \Omega_B(x)} \right|, \min_{1 \leq \mu - N \leq s^+} \left| \frac{d^2_\mu + 1}{d^2_\mu} + \frac{\Omega_A(x)}{b_\mu - \Omega_A(x)} \right| \right\} \gg \kappa^{-1/4} N^{-1/2+\epsilon/2}. \tag{4.13}
\]
Indeed, when 4.13 holds, then the matrices on the left-hand side of 4.12 is non-singular. Note that
\[
\frac{d^2_i + 1}{d^2_i} + \frac{\Omega_B(x)}{a_i - \Omega_B(x)} = \frac{1}{d^2_i} - \frac{a_i}{\Omega_B(x) - a_i} = \frac{a_i}{\Omega_B(\Omega_B^{-1}(\hat{a}_i) - a_i)} - \frac{a_i}{\Omega_B(x) - a_i} = O \left( |\Omega_B(x) - \Omega_B(\Omega_B^{-1}(\hat{a}_i))| \right), \tag{4.14}
\]
where in the last equality we used (i) of Proposition 3.1. The rest of the proof is devoted to control 4.13 using mean value theorem.

First, we have that
\[
|x - \Omega_B^{-1}(\hat{a}_i)| \geq N^{-1/2+\epsilon}\Delta_1(\hat{a}_i), \text{ for all } x \notin I. \tag{4.15}
\]
In fact, when \( i \in \mathcal{O}^{(a)}_e \), 4.11 holds by definition. When \( i \notin \mathcal{O}^{(a)}_e \), by 4.17 and the fact that \( \Omega_B(x) \) is monotone increasing when \( x > E_+ \), we have
\[
\Omega_B^{-1}(\hat{a}_i) - E_+ \lesssim N^{-2/3+2\epsilon} \ll N^{-2/3+3\epsilon}.
\]

26
Now we return to the proof of \([4.13]\). We divide our proof into two cases. If there exists a constant \(c > 0\) such that \(\Omega^{-1}_B(\tilde{a}_i) \notin [x - c\kappa, x + c\kappa]\). Since \(\Omega_B(\cdot)\) is monotonically increasing on \((E_+, \infty)\), we have that

\[
|\Omega_B(x) - \Omega_B(\Omega^{-1}_B(\tilde{a}_i))| \geq |\Omega_B(x) - \Omega_B(x \pm \kappa)| \sim \kappa^{1/2} \gg N^{-1/2+\epsilon/2} \kappa^{-1/4},
\]

where in the second step we used \([4.19]\) and \([5.28]\) with Cauchy’s integral formula when \(x > E_+\). On the other hand, if \(\Omega^{-1}_B(\tilde{a}_i) \in [x - c\kappa, x + c\kappa]\) such that \(\Omega^{-1}_B(\tilde{a}_i) - E_+ \sim \kappa\), here \(c < 1\) is some small constant. By \([4.7]\) and the fact \(\tilde{a}_i - \Omega_B(E_+) \geq N^{-1/3+\epsilon}\), we have that

\[
\Omega^{-1}_B(\tilde{a}_i) - E_+ \sim \Delta_1(\tilde{a}_i)^4 \gg N^{-1/2+\epsilon} \Delta_1(\tilde{a}_i).
\]

Moreover, by \([3.10]\) and \([5.28]\), we conclude that

\[
|\Omega_B(x) - \Omega_B(\Omega^{-1}_B(\tilde{a}_i))| \sim \Delta_1(\tilde{a}_i)^{-2}, \quad \xi \in I_i^{(a)},
\]

where we used \([4.7]\) in the second step. Since \(\Omega_B\) is monotonically increasing on \((E_+, \infty)\), for \(x \notin I_i^{(a)}\), by \([4.15]\) and \([4.17]\), we conclude that

\[
|\Omega_B(x) - \Omega_B(\Omega^{-1}_B(\tilde{a}_i))| \sim \Delta_1(\tilde{a}_i)^{-2}, \quad \xi \in I_i^{(a)}.
\]

The \(d_i^b\) term can be dealt with in the same way, and this completes our proof. \(\square\)

**Steps 2 and 3:** For the ease of discussion, we relabel all the spikes with indices in \(O_e^{(a)} \cup O_e^{(b)}\) as \(\sigma_1, \sigma_2, \ldots, \sigma_r\) and call them \(\epsilon\)-spikes. Further, we assume that they correspond to the classical locations of the outlying eigenvalues as \(x_1 \geq x_2 \geq \cdots \geq x_r\), where some of are determined by \(\Omega^{-1}_A\) while others are given by \(\Omega^{-1}_B\). Correspondingly, all the permissible intervals are relabelled as \(I_i, 1 \leq i \leq r\). We claim that, for a specific \(N\)-independent configuration \(\mathbf{x} \equiv \mathbf{x}(0) := (x_1, \ldots, x_r)\) satisfying \(x_1 > x_2 > \cdots > x_r\), each \(I_i(\mathbf{x})\) contains precisely one eigenvalues of \(\tilde{Q}_1\). This finishes Step 2. For Step 3, we use a continuity argument to generalize the results of Step 2 with configuration \(\mathbf{x}(0)\) to arbitrary \(N\)-dependent configuration and this proves \([4.9]\).

The justification of the above discussion makes use of Rouche’s theorem, \([4.17], [4.10], [4.15]\) and \([5.28]\). We can verbatimly follow the counterparts in Steps 2 and 3 of the proof of Theorem 3.6 of \([20]\) to complete it. We omit further details here.

**Step 4:** In this step, we consider the extremal non-outlier eigenvalues when \(i \notin (O_e^{(a)} \cup O_e^{(b)})\) and prove \([4.11]\). The discussion will use the following eigenvalue interlacing result.

**Lemma 4.4.** Recall the eigenvalues of \(\tilde{Q}_1\) and \(Q_1\) are denoted as \(\{\tilde{\lambda}_i\}\) and \(\{\lambda_i\}\), respectively. Then we have that

\[
\tilde{\lambda}_i \in [\lambda_i, \lambda_i - r - s],
\]

where we adopt the convention that \(\lambda_i = \infty\) if \(i < 1\) and \(\lambda_i = 0\) if \(i > N\).

**Proof.** See \([20]\) Lemma S.3.3). \(\square\)

We first fix a configuration \(\mathbf{x}(0)\) as mentioned earlier. Then by the discussion of Step 2, \([4.10]\) and Lemma \([4.4]\) we can prove \([4.11]\) under the configuration \(\mathbf{x}(0)\). For arbitrary \(N\)-dependent configuration, we again use a continuity argument as mentioned in Step 3. We refer the readers to Step 4 of the proof of Theorem 3.6 of \([20]\). This finishes the proof of Theorem 2.20. \(\square\)
4.2 Eigenvector statistics: proof of Theorem 2.23

In this section, we prove Theorem 2.23. Due to similarity, we only focus on the left singular vectors. The main technical task of this section is to prove Proposition 4.5, which implies Theorem 2.23. Recall the definitions in (1.10) and (2.23).

**Proposition 4.5.** Suppose the assumptions of Theorem 2.23 hold. Then for all \( i, j = 1, 2, \ldots, N \), we have that

\[
\begin{align*}
\left| \langle e_i, P_S e_j \rangle - \delta_{ij} \mathbb{I}(\pi_a(i) \in S) \hat{a}_i \left( \frac{\Omega_B^{-1}(\hat{a}_i)}{\Omega_B^{-1}(\hat{a}_i)} \right) \right| &< \frac{\mathbb{I}(\pi_a(i) \in S, \pi_a(j) \in S)}{\sqrt{N} \sqrt{\Delta_1(\hat{a}_i) \Delta_1(\hat{a}_j)}} + \frac{\mathbb{I}(\pi_a(i) \in S, \pi_b(j) \notin S)}{\sqrt{N} \delta_{\pi_a(i), \pi_b(j)}} \\
+ \frac{1}{N} \left( \frac{1}{\delta_{\pi_a(i)(S)} - \Delta_1(\hat{a}_i)^2} \right) &< \frac{1}{\delta_{\pi_a(j)(S)} - \Delta_1(\hat{a}_j)^2} + \mathbb{I}(\pi_a(j) \in S) + (i \leftrightarrow j),
\end{align*}
\]

where \( (i \leftrightarrow j) \) denotes the same terms but with \( i \) and \( j \) interchanged.

**Proof of Theorem 2.23.** For the right singular vectors, since

\[ v = \sum_{k=1}^{N} \langle e_k, v \rangle e_k = \sum_{k=1}^{N} v_k e_k. \]

Then the results simply follow from Proposition 4.5. The results of the left singular vectors can be obtained similarly.

The rest of the subsection is devoted to the proof of Proposition 4.5. Let \( \omega < \tau_1/2 \) and \( 0 < \epsilon < \min\{\tau_1, \tau_2\}/10 \) be some small positive constants to be chosen later. By Proposition 4.4, Theorems 2.14 and 2.20, we can choose a high probability event \( \Xi_1 = \Xi_1(\epsilon, \omega, \tau_1, \tau_2) \) where the following statements hold.

(i) For all \( z \in \mathcal{D}_{\text{out}}(\omega) := \left\{ E + i\eta : E + N^{-2/3+\omega} \leq E \leq \omega^{-1}, 0 \leq \eta \leq \omega^{-1} \right\} \), we have that

\[
\mathbb{I}(\Xi_1) \left\| \mathbf{U}^*(\mathbf{G}(z) - \Theta(z)) \mathbf{U} \right\| \leq N^{-1/2+\epsilon}(\kappa + \eta)^{-1/4}.
\]

(ii) Recall the notations in (1.10). For all \( 1 \leq i \leq r^+ \) and \( 1 \leq \mu - N \leq s^+ \), we have

\[
\mathbb{I}(\Xi_1) \left| \hat{\lambda}_{\pi_a(i)} - \Omega_B^{-1}(\hat{a}_i) \right| \leq N^{-1/2+\Delta_1(\hat{a}_i)}, \quad \mathbb{I}(\Xi_1) \left| \hat{\lambda}_{\pi_b(\mu)} - \Omega_A^{-1}(\hat{b}_\mu) \right| \leq N^{-1/2+\epsilon} \Delta_1(\hat{b}_\mu).
\]

(iii) For any fixed integer \( \varpi > r + s \) and all \( r^+ + s^+ < i \leq \varpi \), we have that

\[
\mathbb{I} \left| \lambda_1 - E_+ \right| + |\hat{\lambda}_i - E_+| \leq N^{-2/3+\epsilon}.
\]

From now on, we will focus our discussion on the high probability event \( \Xi_1 \) and hence all the discussion will be purely deterministic.

We next provide some contour will be used in our proof. Recall (2.27) and (2.28). Denote

\[
\rho^a_i = c_i \left[ \delta_{\pi_a(i)}(S) \land (\hat{a}_i - \Omega_B(\lambda_+)) \right], \quad \pi_a(i) \in S,
\]

and

\[
\rho^b_\mu = c_\mu \left[ \delta_{\pi_b(\mu)}(S) \land (\hat{b}_\mu - \Omega_A(\lambda_+)) \right], \quad \pi_b(\mu) \in S,
\]

for some sufficiently small constants \( 0 < c_i, c_\mu < 1 \). Define the contour \( \Gamma := \partial C \) as the boundary of the union of the open discs

\[
C := \bigcup_{\pi_a(i) \in S} B_{\rho^a_i}(\hat{a}_i) \cup \bigcup_{\pi_b(\mu) \in S} B_{\rho^b_\mu}(\Omega_B(\Omega_B^{-1}(\hat{b}_\mu))),(4.19)
\]

28
where $B_r(x)$ denotes an open disc of radius $r$ around $x$. In the following lemma, we will show that by choosing sufficiently small $c_1, c_\mu$, we have that: (1). $\partial\Omega_B^{-1}(C)$ is a subset of $\Omega_B^{-1}(C)$ and hence $\Omega_B^{-1}(C)$ holds; (2). $\partial\Omega_B^{-1}(C) = \Omega_B^{-1}(\Gamma)$ only encloses the outliers in $S$. Its proof will be put into Appendix D.

**Lemma 4.6.** Suppose that the assumptions of Theorem 2.20 hold true. Then the set $\Omega_B^{-1}(C)$ lies in the parameter set $(4.16)$ as long as $c_1$’s and $c_\mu$’s are sufficiently small. Moreover, we have that $\{\lambda_a\}_{a \in S} \subset \Omega_B^{-1}(C)$ and all the other eigenvalues lie in the complement of $\Omega_B^{-1}(C)$.

Then we introduce some decompositions. Denote $\hat{H} = H(z)$ as

$$
\hat{H}(z) := PH(z)P = \begin{pmatrix} 0 & z^{1/2} \hat{Y} \cr z^{1/2} \hat{Y}^* & 0 \end{pmatrix}, \\
P = \begin{pmatrix} (1 + D^a)^{1/2} & 0 \\
0 & (1 + D^b)^{1/2} \end{pmatrix},
$$

(4.20)

Correspondingly, we denote $\hat{G}(z) = (\hat{H}(z) - z)^{-1}$. By a discussion similar to (4.1) and the singular value decomposition (SVD) of $Y$, we have that

$$
\hat{G}_{ij} = \frac{1}{\sqrt{\lambda}} \sum_{k=1}^{N} \sqrt{\lambda_k} \hat{u}_k(i) \hat{v}_k(j), \\
\hat{G}_{\mu
u} = \frac{1}{\lambda - z} \sum_{k=1}^{N} \sqrt{\lambda_k} \hat{v}_k(\mu) \hat{v}_k(\nu),
$$

(4.21)

Our starting point is an integral representation. Specifically, by (4.21), Lemma 4.6 and Cauchy’s integral formula, we have that

$$
\langle e_i, P_S e_j \rangle = -\frac{1}{2\pi i} \oint_{\Omega_B^{-1}(\Gamma)} \langle e_i, \hat{G}(z)e_j \rangle dz,
$$

(4.22)

where $e_i$ and $e_j$ are the natural embeddings of $e_i$ and $e_j$ in $\mathbb{C}^{2N}$. Next, we provide an identity of $U^* \hat{G}(z)U$ in terms of $U^*G(z)U$. For the matrices $A, S, B$ and $T$ of conformable dimensions, by Woodbury matrix identity, we have

$$
(A + SBT)^{-1} = A^{-1} - A^{-1}S(B^{-1} + TA^{-1}S)^{-1}TA^{-1},
$$

(4.23)

as long as all the operations are legitimate. Moreover, when $A + B$ is non-singular, we have that

$$
A + (A + B)^{-1}A = B - B(A + B)^{-1}B.
$$

(4.24)

By (4.20), (4.21) and the matrix identities (4.23) and (1.24), we have that

$$
U^* \hat{G}(z)U = U^*P^{-1} (H - z + z(I - P^{-2}))^{-1} P^{-1}U = U^*P^{-1}(G^{-1}(z) + zU^*G(z)U^{-1}P^{-1}U
$$

$$
= U^* P^{-1} \left[ G(z) - zG(z)U \frac{1}{D^{-1} + zU^*G(z)U} U^*G(z)U \right] P^{-1}U
$$

$$
= \tilde{D}^{1/2} \left[ U^*G(z)U - zU^*G(z)U \frac{1}{D^{-1} + zU^*G(z)U} U^*G(z)U \right] \tilde{D}^{1/2}
$$

$$
= \frac{1}{\tilde{D}} \tilde{D}^{1/2} \left[ D - D^{-1} \frac{1}{D^{-1} + zU^*G(z)U} D^{-1} \right] \tilde{D}^{1/2},
$$

(4.25)

where $\tilde{D} = P^{-2}$ and $P$ is defined in (4.20). Then we prove Proposition 4.5. The proof follows from the state strategy as [20, Proposition S.5.5]
**Proof of Proposition 4.25**  Till the end of the proof, for convenience, we set \( d_i^0 = 0 \) when \( i > r \). Denote \( \mathcal{E}(z) = z \mathcal{U}^*(\Theta(z) - G(z)) \mathcal{U} \). Using the resolvent expansion, we obtain that
\[
\frac{1}{\mathcal{D}^{-1} + z \mathcal{U}^* \mathcal{G}(z) \mathcal{U}} = \frac{1}{\mathcal{D}^{-1} + z \mathcal{U}^* \Theta(z) \mathcal{U}} + \frac{1}{\mathcal{D}^{-1} + z \mathcal{U}^* \Theta(z) \mathcal{U}} \mathcal{E} \frac{1}{\mathcal{D}^{-1} + z \mathcal{U}^* \Theta(z) \mathcal{U}}.
\]
Together with (4.22) and (4.25), using the fact that \( \Gamma \) does not enclose any pole of \( G \) (i.e. (4.18)), we have the following decomposition
\[
\langle e_i, \mathcal{P}_S e_j \rangle = \frac{\sqrt{(1 + d_i^0)(1 + d_j^0)}}{d_i^0 d_j^0} (s_0 + s_1 + s_2),
\]
where \( s_0, s_1 \) and \( s_2 \) are defined as
\[
s_0 = \frac{\delta_{ij}}{2\pi i} \oint_{\Omega_B^{-1}(\Gamma)} \frac{1}{(d_i^0)^{-1} + a_i (a_i - \Omega_B(z))^{-1}} \frac{dz}{z},
\]
\[
s_1 = \frac{1}{2\pi i} \oint_{\Omega_B^{-1}(\Gamma)} \frac{\mathcal{E}_{ij}(z)}{((d_i^0)^{-1} + a_i (a_i - \Omega_B(z))^{-1}) ((d_j^0)^{-1} + a_j (a_j - \Omega_B(z))^{-1})} \frac{dz}{z},
\]
\[
s_2 = \frac{1}{2\pi i} \oint_{\Omega_B^{-1}(\Gamma)} \left( \frac{1}{\mathcal{D}^{-1} + z \mathcal{U}^* \Theta(z) \mathcal{U}} \mathcal{E} \frac{1}{\mathcal{D}^{-1} + z \mathcal{U}^* \Theta(z) \mathcal{U}} \right) \frac{dz}{z}.
\]
First, we deal with the term containing \( s_0 \). Using residual theorem, we readily see that
\[
\frac{\sqrt{(1 + d_i^0)(1 + d_j^0)}}{d_i^0 d_j^0} s_0 = \frac{\sqrt{(1 + d_i^0)(1 + d_j^0)}}{d_i^0 d_j^0} \frac{\delta_{ij}}{2\pi i} \oint_{\Gamma} \frac{(\Omega_B^{-1})' (\zeta) a_i - \zeta}{\Omega_B^{-1}(\zeta)} d\zeta = \delta_{ij} \hat{a}_i \frac{(\Omega_B^{-1})'(\hat{a}_i)}{\Omega_B^{-1}(\hat{a}_i)}.
\]
Second, we control the term containing \( s_1 \). For the ease of discussion, we further apply residual theorem to obtain that
\[
s_1 = \frac{d_i^0 d_j^0}{2\pi i} \oint_{\Gamma} \frac{\xi_{ij}(\zeta) \xi_{ij}(\zeta)}{(\zeta - \hat{a}_i)(\zeta - \hat{a}_j)} d\zeta, \quad \xi_{ij}(\zeta) = (\zeta - a_i)(\zeta - a_j)\mathcal{E}_{ij}(\Omega_B^{-1}(\zeta)) \frac{\Omega_B^{-1}(\zeta)}{\Omega_B^{-1}(\hat{a}_i)}. \tag{4.26}
\]
To bound \( \xi_{ij}(\zeta) \), we first prepare some useful estimates. When \( c_i \)'s and \( c_j \)'s are sufficiently small, by a discussion similar to (4.17), we have that for \( \zeta \in \Gamma \),
\[
|\Omega_B^{-1}(\zeta) - E_+| \sim |\zeta - \Omega_B(E_+)|^2. \tag{4.27}
\]
Moreover, let \( z_b = \Omega_B^{-1}(\zeta) \), by Cauchy’s differentiation formula, we obtain that
\[
\Omega_B'(z_b) - \Omega_B'(z_b) = \frac{1}{2\pi i} \oint_{\mathcal{C}_b} \frac{\Omega_B(\xi) - \Omega_B(\xi)}{(\xi - z_b)^2} d\xi, \tag{4.28}
\]
where \( \mathcal{C}_b \) is the disc of radius \( |z_b - E_+|/2 \) centered at \( z_b \). Here we used the facts that both \( \Omega_B \) and \( \Omega_\beta \) are holomorphic on \( \Omega_B^{-1}(\Gamma) \). Together with (3.22) and (3.19), using residual theorem, we readily see that for some constant \( C > 0 \)
\[
\Omega_B'(z_b) \sim C N^{-1/2 + \epsilon} |z_b - E_+|^{-1} + |z_b - E_+|^{-1/2} \sim |z_b - E_+|^{-1/2},
\]
30
where in the last step we used \( \ref{eq:4.27} \) such that \( |z_b - E_+| \geq C N^{-2/3+\epsilon} \). Consequently, using implicit differentiation, we conclude that

\[
(\Omega_B^{-1})'(|\zeta|) \sim \frac{1}{|z_b - E_+|^{1/2}} \sim |\zeta - \Omega_B(E_+)|,
\]

where in the last step we used \( \ref{eq:4.27} \). With the above preparation, we now proceed to control \( \xi_{ij}(\zeta) \) and \( s_1 \).

By \( \ref{eq:4.17}, \ref{eq:4.27} \) and \( \ref{eq:4.29} \), it is easy to see that for \( \zeta \in \Gamma \)

\[
|\xi_{ij}(\zeta)| \lesssim N^{-1/2+\epsilon}|\zeta - \Omega_B(E_+)|^{1/2}.
\]

Together with a discussion similar to \( \ref{eq:4.28} \), we see that

\[
|s_1| \leq C N^{-1/2+\epsilon} \|\frac{\xi_{ij}(\zeta)}{a_i - a_j}\| \lesssim C \sqrt{\Delta_1(a_i)\Delta_1(a_j)},
\]

where we used \( \ref{eq:4.31} \) in the last step. On the other hand, if \( \hat{a}_i = \hat{a}_j \), we can obtain the same bound using residual theorem. In the second case when \( \pi_a(i) \in S \) and \( \pi_b(j) \notin S \), we conclude from \( \ref{eq:4.30} \) that

\[
|s_1| \leq C \Delta_1(\hat{a}_i) N^{-1/2+\epsilon}.
\]

Similarly, we can estimate \( s_1 \) when \( \pi_a(i) \notin S \) and \( \pi_b(j) \in S \). Finally, when both \( \pi_a(i) \notin S \) and \( \pi_b(j) \notin S \), we have \( s_1 = 0 \) by residual theorem. This completes the estimation regarding \( s_1 \).

We then estimate \( s_2 \), which relies on some crucial estimates on the contour. We decompose the contour into

\[
\Gamma = \bigcup_{\pi_a(i) \in S} \Gamma_i \cup \bigcup_{\pi_b(\mu) \in S} \Gamma_\mu, \quad \Gamma_i := \Gamma \cap \partial B_{\rho_a}^{h_i}(\hat{a}_i), \quad \Gamma_\mu := \Gamma \cap \partial B_{\rho_b}^{\nu}(\Omega_B(\Omega^{-1}_A(\hat{b}_\mu))).
\]

The following lemma is the key technical input for the estimation of \( s_2 \). Its proof will be given in Appendix \ref{app:B}.

**Lemma 4.7.** For any \( \pi_a(i) \in S, 1 \leq j \leq r, 1 \leq \nu - N \leq s \) and \( \zeta \in \partial B_{\rho_a}^{h_i}(\hat{a}_i) \), we have that

\[
|\zeta - \hat{a}_j| \sim \rho_i^a + \delta_{\pi_a(i),\pi_a(j)},
\]

and

\[
\Omega_A(\Omega^{-1}_B(\zeta)) - \hat{b}_\nu \sim \rho_i^a + \delta_{\pi_a(i),\pi_a(\nu)}.
\]

For any \( \pi_b(\mu) \in S, 1 \leq j \leq r, 1 \leq \mu - N \leq s \) and \( \zeta \in \partial B_{\rho_b}^{\nu}(\Omega_B(\Omega^{-1}_A(\hat{b}_\mu))) \), we have

\[
|\zeta - \hat{a}_i| \sim \rho_i^b + \delta_{\pi_b(\mu),\pi_a(j)},
\]

and

\[
\Omega_A(\Omega^{-1}_B(\zeta)) - \hat{b}_\nu \sim \rho_i^b + \delta_{\pi_b(\mu),\pi_a(\nu)}.
\]

For the estimation of \( s_2 \), by \( \ref{eq:4.17}, \ref{eq:4.30}, \ref{eq:4.27} \) and (i) of Proposition \ref{prop:3.1} we have that

\[
|s_2| \leq C \int_{\Gamma} \frac{N^{-1/2+\epsilon}}{|\zeta - \hat{a}_i| |\zeta - \hat{a}_j|} \left| \frac{(\Omega_B^{-1})'(\zeta)}{|\zeta - \Omega_B(E_+)|} \right| \left\| \mathcal{D}^{-1} + \Omega_B^{-1}(\zeta) U^* G(\Omega_B^{-1}(\zeta)) U \right\| |d\zeta|,
\]

\[
\leq C \int_{\Gamma} \frac{1}{|\zeta - \hat{a}_i| |\zeta - \hat{a}_j|} \left| \frac{d(\zeta) - \|E(\Omega_B^{-1}(\zeta))\|}{\mathcal{E}(\Omega_B^{-1}(\zeta))} \right| |d\zeta|,
\]

\(31\)
where \( \mathfrak{d}(\zeta) \) is defined as
\[
\mathfrak{d}(\zeta) := \left( \min_{1 \leq j \leq r} |\hat{a}_j - \Omega_B(\zeta)| \right) \wedge \left( \min_{1 \leq \mu - N \leq s} |\hat{\beta}_\mu - \Omega_A(\Omega_B^{-1}(\zeta))| \right).
\]
We mention that \( \mathfrak{d}(\zeta) \) can be bounded using Lemma 4.7. Moreover, by (4.17) and (4.27), for some constant \( C > 0 \), we can bound
\[
\|\mathcal{E}(\Omega_B^{-1}(\zeta))\| \leq C \sqrt{rs} N^{-1/2+\epsilon} |\zeta - \Omega_B(E_+)|^{-1/2}.
\]
(4.35)
Recall that both \( r \) and \( s \) are bounded. Together with Lemma 4.7 and the fact \( \epsilon < \tau_2 \), we obtain that
\[
\|\mathcal{E}(\Omega_B^{-1}(\zeta))\| \ll (\hat{a}_i - \Omega_B(E_+))^{-1/2} N^{-1/2+\tau_2} \lesssim \begin{cases} 
\rho^a_i \lesssim \mathfrak{d}(\zeta), & \text{for } \zeta \in \Gamma_i \\
\rho^b_\mu \lesssim \mathfrak{d}(\zeta), & \text{for } \zeta \in \Gamma_\mu,
\end{cases}
\]
where we used (4.35) and Assumption 2.22. Based on the above estimates, we arrive at
\[
\frac{1}{\mathfrak{d}(\zeta) - \|\mathcal{E}(\Omega_B^{-1}(\zeta))\|} \lesssim \begin{cases} 
(\rho^a_i)^{-1}, & \text{for } \zeta \in \Gamma_i \\
(\rho^b_\mu)^{-1}, & \text{for } \zeta \in \Gamma_\mu.
\end{cases}
\]
(4.36)
Now we proceed to control \( s_2 \). Decomposing the integral contour in (4.34) as in (4.32), using (4.36) and Lemma 4.7, and recalling that the length of \( \Gamma_i \) (or \( \Gamma_\mu \)) is at most \( 2\pi \rho^a_i \) (or \( 2\pi \rho^b_\mu \)), we get that for some constant \( C > 0 \),
\[
|s_2| \leq C \sum_{\pi_a(k) \in S} \frac{N^{-1+2\epsilon}}{(\rho^a_k + \delta_{\pi_a(k),\pi_a(i)})^2} + C \sum_{\pi_b(\mu) \in S} \frac{N^{-1+2\epsilon}}{(\rho^b_\mu + \delta_{\pi_b(\mu),\pi_b(i)})^2}.
\]
(4.37)
Now we bound the terms of the right-hand side of (4.37) using Cauchy-Schwarz inequality. For \( \pi_a(i) \notin S \), we have
\[
\sum_{\pi_a(k) \in S} \frac{1}{(\rho^a_k + \delta_{\pi_a(k),\pi_a(i)})^2} \leq \sum_{\pi_a(k) \in S} \frac{1}{(\delta_{\pi_a(k),\pi_a(i)})^2} \leq \frac{C}{\delta_{\pi_a(i)}(S)^2}.
\]
For \( \pi_a(i) \in S \), we have \( \rho^a_k + \delta_{\pi_a(k),\pi_a(i)} \gtrsim \rho^a_i \) for \( \pi_a(k) \in S \), and \( \rho^b_\mu + \delta_{\pi_b(\mu),\pi_b(i)} \gtrsim \rho^b_i \) for \( \pi_b(\mu) \in S \). Then we have for some constant \( C > 0 \)
\[
\sum_{\pi_a(k) \in S} \frac{1}{(\rho^a_k + \delta_{\pi_a(k),\pi_a(i)})^2} \leq \frac{C}{(\rho^a_i)^2} \leq \frac{C}{\delta_{\pi_a(i)}(S)^2} + \frac{C}{\Delta_1(\hat{a}_i)^2}.
\]
Plugging the above two estimates into (4.37), we get that
\[
|s_2| \leq CN^{-1+2\epsilon} \left( \frac{1}{\delta_{\pi_a(i)}(S)} + \frac{I(\pi_a(i) \in S)}{\Delta_1(\hat{a}_i)^2} \right) \left( \frac{1}{\delta_{\pi_a(j)}(S)} + \frac{I(\pi_a(j) \in S)}{\Delta_1(\hat{a}_j)^2} \right).
\]
So far, we have proved Proposition 4.5 for \( 1 \leq i, j \leq r \) since \( \epsilon \) can be arbitrarily small.
Finally, the general case can be dealt with easily. For general \( i, j \in \{1, \cdots, N\} \), we define \( \mathcal{R} := \{1, \cdots, r\} \cup \{i, j\} \). Then we define a perturbed model as
\[
\hat{A} = A \left( I + \hat{D}^a \right), \quad \hat{D}^a = \text{diag}(d^a_k)_{k \in \mathcal{R}},
\]
where for some \( \bar{c} > 0 \),
\[
d^a_k := \begin{cases} 
\bar{d}^a_k, & \text{if } 1 \leq k \leq r \\
\bar{c}, & \text{if } k \in \mathcal{R} \text{ and } k > r.
\end{cases}
\]

32
Then all the previous proof goes through for the perturbed model as long as we replace the $U$ and $D$ in (4.2) with
$$
\hat{U} = \begin{pmatrix} E_{r+2} & 0 \\ 0 & E_{s} \end{pmatrix}, \quad \hat{D} = \begin{pmatrix} \hat{D}^a(D^a + 1)^{-1} & 0 \\ 0 & \hat{D}^b(D^b + 1)^{-1} \end{pmatrix}.
$$
(4.38)

Note that in the proof, only the upper bound on the $d_{ik}$’s were used. Moreover, the proof does not depend on the fact that $\hat{a}_i$ or $\hat{a}_j$ satisfy (2.22) (we only need the indices in $S$ to satisfy Assumption 2.22). By taking $\tilde{\epsilon} \downarrow 0$ and using continuity, we get that Proposition 4.5 holds for general $i, j \in \{1, \cdots, N\}$. □

5 Proof of Theorems 2.14 and 2.15

In this section, we prove the spectral rigidity near the upper edge, i.e., Theorem 2.14 and the complete delocalization of the singular vectors, i.e., Theorem 2.15.

We start with Theorem 2.14. For the first part of the results, i.e., the eigenvalues are close to the quantiles $\gamma^*_i$, the proof follows from a standard argument established in [33], by translating the closeness of the resolvent into the closeness of the eigenvalues and the quantiles of $\mu_A \otimes \mu_B$. Since the proof strategy is rather standard, we only sketch the key inputs and steps. Overall, the proof rely on the following two important inputs:

1. For the largest eigenvalue $\lambda_1$, we have
   $$|\lambda_1 - \gamma^*_1| \ll N^{-2/3}.
   $$
   (5.1)

2. Given sufficiently small constant $\zeta > 0$, we have
   $$\sup_{x \leq E_+ + \zeta} |\mu_H((\infty, x]) - \mu_A \otimes \mu_B((\infty, x]| \ll N^{-1}.
   $$
   (5.2)

The proof of the first part of Theorem 2.14 follow from the above two claims and the square root behavior of $\mu_A \otimes \mu_B$ established in (ii) of Proposition 3.1. We mention that the second technical input is from an argument regardless of the underlying random matrix models once we have established the local laws in Theorem 2.12. In detail, it follows from a standard application of the Helffer-Sjöstrand formula on $D_\tau(\eta_L, \eta_U)$; see Section C and the arguments in Section 8 of [10] for details. In what follows, we focus our discussion on the justification of the first claim since it usually depends on a case by case checking.

For the second part of Theorem 2.14 when we replace $\gamma^*_i$ with $\gamma_i$, we will bound the differences between $\gamma^*_i$ and $\gamma_i$.

Proof of Theorem 2.14. First, we prove the first part by proving (5.1) by contradiction. It relies on the following improve estimates, whose proof can be found in Appendix D. Denote $\tilde{D}$ as
$$
\tilde{D} := \{z = E + i\eta \in D_\tau(\eta_L, \eta_U) : \sqrt{\kappa + \eta} > \frac{N^{2/3}}{\sqrt{\eta}} \text{ and } E > E_+\}.
$$
(5.3)

Lemma 5.1. Under the assumptions of Theorem 2.14, we have the following holds uniformly in $z \in \tilde{D}$
$$
\Lambda(z) \ll \frac{1}{N(\kappa + \eta)^{1/2}} + \frac{1}{\sqrt{\kappa + \eta} (N\eta)^{3/2}}.
$$
(5.4)

Recall (2.11). For any (small) constant $\epsilon > 0$, we define the following line segment
$$
\tilde{D}(\epsilon) := \{z = E + i\eta : E_+ + N^{-2/3+6\epsilon} \leq E < \tau^{-1}, \ \eta = N^{-2/3+\epsilon}\},
$$
where $\tau > 0$ is some small fixed constant. Clearly, we have that $\mathcal{D}(\epsilon) \subset \mathcal{D}_\tau$. By \textbf{(4.1)}, we readily obtain that $\Lambda \lesssim \frac{N^{-\epsilon}}{N\eta}$ holds uniformly in $z \in \mathcal{D}(\epsilon)$. Together with $(7.45)$, we see that for $z \in \mathcal{D}(\epsilon)$ uniformly

$$|m_H(z) - m_{\mu\alpha}\mu\beta(z)| \lesssim \frac{N^{-\epsilon}}{N\eta}.$$ 

Together with (ii) of Proposition \textbf{3.1} we arrive at

$$\text{Im} \ m_H(z) \sim \frac{N^{-\epsilon}}{N\eta}, \tag{5.5}$$

holds uniformly in $z \in \mathcal{D}(\epsilon)$. By \textbf{(2.11)}, it suffices to prove that with high probability $\lambda_1 \leq E_+ + N^{-2/3+6\epsilon}$ in order to prove \textbf{(5.1)}. We now justify it by contradiction. Suppose that $\lambda_1 > E_+ + N^{-2/3+6\epsilon}$, then for any $\eta > 0$ and by the definition of Stieltjes transform, we see that

$$\sup_{E > E_+ + N^{-2/3+6\epsilon}} \text{Im} \ m_H(E + i\eta) = \sup_{E > E_+ + N^{-2/3+6\epsilon}} \frac{1}{N} \sum_{k=1}^{N} \left( \lambda_k - E \right)^2 + \eta^2 \geq \frac{1}{N\eta},$$

which is a contradiction to \textbf{(5.5)}. Therefore, we have proved \textbf{(5.1)}.

Next, for the second part of the results, we first claim the following estimate.

\textbf{Lemma 5.2.} Suppose the assumptions of Theorem \textbf{2.14} hold. Then for some sufficiently large $N_0 \equiv N_0(\epsilon, c)$, when $N \geq N_0$, we have

$$|\gamma_i - \gamma_i^*| \leq j^{-1/3}N^{-2/3+\epsilon}, \quad 1 \leq j \leq cN. \tag{5.6}$$

It is easy to see that the second part of the results follow from \textbf{(5.6)} and the first part of the results. The rest of the proof is devoted to proving \textbf{(5.6)}. Similar to the previous discussion of the first part of the results, we will follow the proof strategy of \cite[Lemma 3.14]{[7]} and focus on establishing results analogous to \textbf{(5.1)} and \textbf{(5.2)}. More specifically, we prove the following results:

$$|\gamma_1 - \gamma_1^*| \lesssim N^{-2/3}, \tag{5.7}$$

$$\sup_{x \leq E_+ + \xi} |\mu_{\alpha} \mu_{\beta}((-\infty, x]) - \mu_{\alpha} \mu_{\beta}(-\infty, x)]| \leq CN^{-1+\epsilon}. \tag{5.8}$$

\textbf{Proof.} We first show \textbf{(5.8)}. By the continuity of the free multiplicative convolution, i.e. Proposition 4.14 of \cite{[11]}, and (iv) of Assumption \textbf{2.4} we obtain that

$$d_L(\mu_{\alpha} \mu_{\beta}, \mu_{\alpha} \mu_{\beta}) \leq d_L(\mu_{\alpha}, \mu_{\alpha}) + d_L(\mu_{\beta}, \mu_{\beta}) \leq N^{-1+\epsilon}.$$ 

Then we can conclude that for some constant $C > 0$,

$$\sup_{x \leq E_+ + \xi} |\mu_{\alpha} \mu_{\beta}((-\infty, x]) - \mu_{\alpha} \mu_{\beta}(-\infty, x)]| \leq CN^{-1+\epsilon},$$

where we used the definition of the Levy distance and \textbf{(2.11)}. Then we show the counterpart of \textbf{(5.1)}. By \textbf{(3.23)}, the definition of subordination function \textbf{(2.8)} and the inverse formula of Stieltjes transform, we obtain that

$$\gamma_1 \leq E_+ + N^{-1+\xi},$$

where we recall \textbf{(3.21)} that $\xi > 1$ is some fixed constant. Together with (ii) of Proposition \textbf{3.1} by definition, we obtain that

$$\frac{j}{N} = \int_{\gamma_j}^{\infty} d\mu_{\alpha} \mu_{\beta}(x) = \int_{\gamma_j}^{E_+ + N^{-1+\xi}} d\mu_{\alpha} \mu_{\beta}(x) \leq C \int_{\gamma_j}^{E_+ + N^{-1+\xi}} \text{Im} \ m_{\alpha}\mu_{\beta}(x + iN^{-1+\xi}) dx,$$
\[ \leq C \int_{\gamma_j}^{E_+ + N^{-1+\varepsilon}} \left[ |E_+ - x| + N^{-1+\varepsilon} \right]^{1/2} dx \]

\[ \leq C |E_+ - \gamma_j^*|^3/2 + CN^{-1+\varepsilon}|E_+ - \gamma_j^*|, \]

where in the second step we used the inversion formula for the Stieltjes transform. This yields that for some constant \( c > 0 \),

\[ |\gamma_j^* - E_+| \geq cN^{-2/3}j^{2/3}. \]

Together with the fact that \( E_+ - \gamma_j \sim N^{-2/3}j^{2/3} \) by (2.13), we conclude that

\[ \gamma_j^* \leq E_+ - cN^{-2/3+\varepsilon}, \]

whenever \( \gamma_j \leq E_+ - cN^{-2/3+\varepsilon} \). Similarly, we can show that \( \gamma_j^* \geq E_+ +-CN^{-2/3+\varepsilon} \) whenever \( \gamma_j \leq E_+ - CN^{-2/3+\varepsilon} \). Based on the above arguments, we conclude that for some constant \( C_1 > 0 \),

\[ |\gamma_j - \gamma_j^*| \leq E_+ - \gamma_j + |E_+ - \gamma_j^*| \leq C_1 N^{-2/3+\varepsilon}, \]

whenever \( E_+ - N^{-2/3+\varepsilon} \leq \gamma_j \leq E_+ \).

This establishes (5.7) and hence completes the proof.

Then we prove Theorem 2.15.

**Proof of Theorem 2.15.** We focus our discussion on the left singular vectors. By spectral decomposition, we have that

\[ \sum_{k=1}^{N} \frac{\eta|u_k(i)|^2}{(\lambda_k - E)^2 + \eta^2} = \text{Im} \tilde{G}_{ii}(z), \quad z = E + i\eta. \] (5.9)

For \( 1 \leq k_0 \leq cN \), let \( z_0 = \lambda_k + i\eta_0 \), where \( \eta_0 = N^{-1+c_0}, c_0 > \gamma \) is a small constant. Inserting \( z = z_0 \) into (5.9), by Theorem 2.12 and (i) of Proposition 3.1, we immediately see that

\[ |u_k(i)|^2 < \eta_0 \lesssim N^{-1}. \]

This concludes our proof.

## 6 Pointwise local laws

In this section, we provide several controls for the entries of the resolvents for each fixed spectral parameter \( z \) under certain assumption (c.f. Assumption 6.2 and the assumption (6.81)). The general and optimal case, i.e., Theorem 2.12 can be proved by removing Assumption 6.2 and (6.81) and using a standard dynamic bootstrapping procedure in next section. Moreover, we prove an optimal estimate (up to some \( N^\varepsilon \) error for arbitrarily small \( \varepsilon > 0 \)) for some functionals of the resolvents, which are the key ingredients for the bootstrapping procedure. Till the end of the paper, we focus our discussion on the Haar unitary matrix and only point out the main technical difference with the Haar orthogonal random matrices along the proof.

### 6.1 Main tools

In this section, we prepare some important identities for our later calculation. We first introduce some analytic functions, which are good approximations for the subordination functions and the diagonal entries of the Green functions.
Definition 6.1 (Approximate subordination functions). For $z \in \mathbb{C} \setminus \mathbb{R}_+$, we define
\[
\Omega_A^r = \Omega_A^r(z) := \frac{z \text{tr} \bar{B}G}{1 + z \text{tr} G}, \quad \Omega_B^r = \Omega_B^r(z) := \frac{z \text{tr} AG}{1 + z \text{tr} G} = \frac{z \text{tr} \bar{G}A}{1 + z \text{tr} \bar{G}}.
\]

Next, we summarize some resolvent identities which will be used in the proof of local laws. Recall (6.4). Using the trivial relations $G(H - z) = I$ and $(H - z)\bar{G} = I$, it is not hard to see that
\[
(HG)_{ii} - zG_{ii} = a_{ii}(\bar{B}G)_{ii} - zG_{ii} = 1, \quad (\bar{G}H)_{ii} - zG_{ii} = b_{ii}(\bar{G}A)_{ii} - zG_{ii} = 1.
\]
Moreover, using the fact
\[
G = A^{1/2}\bar{G}A^{-1/2}, \quad \bar{G} = B^{-1/2}\bar{G}B^{1/2},
\]
we readily obtain that
\[
G^*G = A^{-1/2}\bar{G}^*A^{-1/2}, \quad \bar{G}^*G = B^{1/2}\bar{G}B^{-1/2}.
\]

Our proof makes use of the partial randomness decomposition for Haar unitary matrix. This technique has been employed in studying the addition of random matrices in [3, 4, 5, 7]. Let $U$ be the $(N \times N)$ Haar unitary random matrix. For all $i \in [1, N]$, define $v_i := Ue_i$ as the $i$-th column vector of $U$ and $\theta_i$ as the argument of $e_i^*v_i$. Following [17], we denote
\[
U^{(i)} := -e^{-i\theta_i}R_iU, \quad \text{where} \quad R_i := I - r_ir_i^* \quad r_i := \sqrt{2}\frac{e_i^* + e^{-i\theta_i}v_i}{\|e_i + e^{-i\theta_i}v_i\|_2}.
\]
Since $\|r_i\|_2^2 = 2$, we have that $R_i$ is a Householder reflection. Consequently, we have that $R_i^2 = R_i$ and $R_i^2 = I$. Furthermore, it is elementary to see that $U^{(i)}e_i = e_i$, and $e_i^*U^{(i)} = e_i^*$. This implies that $U^{(i)}$ is a unitary block-diagonal matrix, that is to say, $U^{(i)} = 1$ and the $(i, i)$-matrix minor of $U^{(i)}$ is Haar distributed on $\mathcal{U}(N - 1)$ and $v_i$ is uniformly distributed on the $N - 1$ unit sphere. We next introduce some notations which will be frequently used throughout the proof. Denote
\[
\tilde{B}^{(i)} := U^{(i)}B(U^{(i)})^*.
\]
Since $v_i$ is uniformly distributed on the unit sphere $S_{N-1}^N$, we can find a Gaussian vector $\tilde{g}_i \sim \mathcal{N}_\mathbb{C}(0, N^{-1}I_N)$ such that
\[
v_i = \frac{\tilde{g}_i}{\|\tilde{g}_i\|}.
\]
Armed with the above Gaussian vector, we define
\[
g_i := e^{-i\theta_i}\tilde{g}_i, \quad h_i := \frac{g_i}{\|g_i\|} = e^{-i\theta_i}v_i, \quad \ell_i := \frac{\sqrt{2}}{\|e_i + h_i\|_2}, \quad \tilde{g}_i := g_i - g_i e_i, \quad \tilde{h}_i := h_i - h_i e_i.
\]
Recall (6.4). We have that $r_i = \ell_i(e_i + h_i)$. In addition, we have
\[
R_i e_i = -h_i \quad \text{and} \quad R_i h_i = -e_i.
\]
This implies that
\[
h_i^*\tilde{B}^{(i)} R_i = -e_i^*\tilde{B}, \quad e_i^*\tilde{B}^{(i)} R_i = -h_i^*\tilde{B} = -b_i h_i^*.
\]
In this paper, we focus on the discussion of Haar unitary matrix. For Haar orthogonal random matrix on $O(N)$, the only difference lies in the partial randomness decomposition. In fact, we can decompose $U$ in the same way as in (6.4), except that the factor $e^{-i\theta_i}$ in (6.4) should be replaced by $\text{sgn}(e_i^*v_i)$. We refer the readers to [5] Appendix A for more details.
In the rest of this section, we introduce some useful decompositions. Without loss of generality, we assume that both $A$ and $B$ are normalized such that $\text{tr} A = \text{tr} B = 1$, where we recall the definition (2.1). Recall that we always use $z = E + i\eta$. Throughout the paper, we use the following control parameters

$$\Psi \equiv \Psi(z) := \sqrt{\frac{1}{N\eta}}, \quad \Pi \equiv \Pi(z) := \sqrt{\frac{\text{Im} \mu_\infty \Sigma_{\infty A}(z)}{N\eta}}, \quad \Pi_i \equiv \Pi_i(z) := \sqrt{\frac{\text{Im} G_{ii}(z) + \text{Im} \gamma_{ii}(z)}{N\eta}}. \quad (6.10)$$

Next, we show some important decomposition regarding the approximate subordination functions. By (6.1) and Definition 6.1, we observe that

$$\text{tr} S_i = \text{tr}(\tilde{B}G)_{ii} = \frac{a_i}{a_i - \Omega_B^c} \left( (z \text{tr} G + 1)(a_i - \Omega_B^c)(\tilde{B}G)_{ii} - (z \text{tr} G + 1) \right)$$

$$= \frac{a_i}{(1 + z \text{tr} G)(a_i - \Omega_B^c)} \left( a_i(z \text{tr} G + 1)(\tilde{B}G)_{ii} - z \text{tr}(AG)(\tilde{B}G)_{ii} - (z \text{tr} G + 1) \right)$$

$$= \frac{a_i z}{(1 + z \text{tr} G)(a_i - \Omega_B^c)} (G_{ii} \text{tr}(A\tilde{B}G) - \text{tr}(GA)(\tilde{B}G)_{ii}). \quad (6.11)$$

Consequently, by Proposition 5.1 to prove (2.10), it suffices to show that $\Omega_B^c$ is close to $\Omega_B$ and control the following quantity

$$Q_i := G_{ii} \text{tr}(A\tilde{B}G) - \text{tr}(GA)(\tilde{B}G)_{ii}. \quad (6.12)$$

In what follows, we present detailed decomposition of $Q_i$. In particular, we discuss the decomposition of $(\tilde{B}G)_{ii}$ using the partial randomness decomposition. Our goal is to explore the independence structure. Recall (6.5) and (6.7). For convenience, we introduce the notations

$$S_i := h_i^* \tilde{B}^{(i)} G e_i, \quad \hat{S}_i := h_i^* \tilde{B}^{(i)} G e_i, \quad T_i := h_i^* G e_i = e^{i\theta_i} e_i^* U^* G e_i, \quad \text{and} \quad \hat{T}_i := h_i^* G e_i, \quad (6.13)$$

where we used $(e^{i\theta})^* = e^{i\theta}$. By the construction of $U^{(i)}$ in (6.4) and (6.8), we find that

$$(\tilde{B}G)_{ii} = -h_i^* \tilde{B}^{(i)} R_i G e_i.$$ 

Using the definition of $R_i$ in (6.4), we can further write

$$(\tilde{B}G)_{ii} = -h_i^* \tilde{B}^{(i)} G e_i + \ell_i^2 h_i^* \tilde{B}^{(i)} (e_i + h_i)(e_i + h_i)^* G e_i,$$

where we recall that $r_i = l_i(e_i + h_i)$. With the notations in (6.13), we can further write that

$$(\tilde{B}G)_{ii} = -S_i + \ell_i^2 (h_i^* \tilde{B}^{(i)} e_i + h_i^* \tilde{B}^{(i)} h_i)(G_{ii} + T_i).$$

Further, since $R_i$ is a projection satisfying (6.8), we have that

$$(\tilde{B}G)_{ii} = -S_i + \ell_i^2 (-b_i h_i^* R_i h_i + h_i^* \tilde{B}^{(i)} h_i)(G_{ii} + T_i)$$

$$= -S_i + \ell_i^2 (b_i h_{ii} + h_i^* \tilde{B}^{(i)} h_i)(G_{ii} + T_i) \quad (6.14)$$

Note that $\tilde{B}^{(i)}$ is independent of $h_i$. We will see later in our proof (e.g. (6.11)), the discussion boils down to control $S_i$ and $T_i$. The main idea is to employ the technique of the integration by parts with respect to the coordinates of $h_i$. Through the calculation, we will frequently use the following terms

$$P_i := Q_i + (G_{ii} + T_i) Y,$$

$$K_i := T_i + \text{tr}(GA)(b_i T_i + (\tilde{B}G)_{ii}) - \text{tr}(GAB)(G_{ii} + T_i), \quad (6.15)$$

37
where
\[
\Upsilon := (\text{tr}(GAB))^2 - \text{tr}(GA) \text{tr}(BGAB) - \text{tr}(GAB) + \text{tr}(GA). \tag{6.16}
\]

Using the facts \(GAB = A\hat{B}G = zG + I\) and \(\text{tr} B = \text{tr} \hat{B} = 1\), we have
\[
\Upsilon = \text{tr}(A\hat{B}G)(\text{tr}(zG + I) - 1) - \text{tr}(GA)(\text{tr}(\hat{B}(zG + 1)) - \text{tr} \hat{B})
= z \left( \text{tr}(G) \text{tr}(A\hat{B}G) - \text{tr}(GA) \text{tr}(\hat{B}G) \right)
= z \sum_i Q_i. \tag{6.17}
\]

Moreover, we that
\[
\Omega_\hat{A}^2 \Omega_B - zM_{\mu_H}(z) = \left( \frac{z^2}{(1 + zm_{\mu_H}(z))^2} \right) (\text{tr}(GA) \text{tr}(\hat{B}G) - \text{tr}(G) \text{tr}(A\hat{B}G))
= - \frac{z}{(1 + zm_{\mu_H}(z))^2} \Upsilon(z), \tag{6.18}
\]
where in the first equality we used \(A\hat{B}G = zG + I\).

The above discussion focuses on the diagonal entries of the resolvents. Similar arguments hold for the off-diagonal terms by considering the decomposition of \((BG)_{ij}\) using the strategy similar to \(6.14\). In what follows, we introduce the counterparts of the diagonal entries. Corresponding to \(6.11\), we denote
\[
Q_{ij} := \text{tr}(GA)(\hat{B} G)_{ij} - G_{ij} \text{tr}(A\hat{B}G). \tag{6.19}
\]

Moreover, using the fact \(G_{ij} = a_i (BG)_{ij}/z\), which follows from \(zG + I = A\hat{B}G\), we see that
\[
Q_{ij} = \left( \frac{z}{a_i} \text{tr}(GA) - \text{tr}(A\hat{B}G) \right) G_{ij} = \frac{\text{tr}(GAB)}{a_i} (\Omega_B^2 - a_i) G_{ij}. \tag{6.20}
\]

In this sense, it suffices to bound \(Q_{ij}\) in order to prove \(6.19\). Similar to the discussion for the diagonal entries, we will need the following quantities
\[
S_{ij} := \hat{h}_i^* \hat{B}^{(i)} G e_j, \quad \hat{S}_{ij} := \hat{h}_i^* \hat{B}^{(i)} G e_j, \tag{6.21}
T_{ij} := \hat{h}_i^* G e_j, \quad \hat{T}_{ij} := \hat{h}_i^* G e_j,
\]
\[
P_{ij} := Q_{ij} + (G_{ij} + T_{ij}) \Upsilon, \quad K_{ij} = T_{ij} + \text{tr}(GA)(b_i T_{ij} + (BG)_{ij}) - \text{tr}(GAB)(G_{ij} + T_{ij}).
\]

### 6.2 Pointwise local laws

In this section, we prove \(2.10\) and \(2.14\) of Theorem \(2.12\) for each fixed spectral parameter \(z\). Due to similarly, we focus our discussion on the diagonal entries \(2.16\) and only briefly discuss the results of the off-diagonal entries \(2.17\). We start with introducing some notations. Denote
\[
\Lambda_{d, i} := \left| zG_{ii} + 1 - \frac{a_i}{a_i - \Omega_B} \right|, \quad \Lambda_d := \max_i \Lambda_{d, i}, \quad \Lambda_T := \max_i |T_i|, \tag{6.22}
\]
Similarly, we can define \(\Lambda_{d, i}^e\) and \(\Lambda_d^e\) by replacing \(\Omega_B\) with its approximate \(\Omega_B^e\). Moreover, we define that
\[
\bar{\Lambda}_{d, i} := \left| zG_{ii} + 1 - \frac{b_i}{b_i - \Omega_A} \right|, \quad \bar{\Lambda}_d := \max_i \bar{\Lambda}_{d, i}, \quad \bar{\Lambda}_T := \max_i |\bar{e}_i U G e_i|.
\]
We first introduce the following assumption.
Assumption 6.2. Recall $\gamma > 0$ is a small constant defined in (2.15). Fix $z \in \mathcal{D}_{\tau}(\eta_L, \eta_U)$ defined in (2.14), we suppose the followings hold true

\[ \Lambda_d(z) \prec N^{-\gamma/4}, \quad \tilde{\Lambda}_d(z) \prec N^{-\gamma/4}, \quad \Lambda_T \prec 1, \quad \tilde{\Lambda}_T \prec 1. \]  

(6.23)

In Section 6, for the ease of statements, we prove all the results under Assumption 6.2. As we will see later in Section 8.1, Assumption 6.2 can be easily removed by establishing the weak local laws; see Proposition 8.1 for more details. The main result of this section is the following proposition.

Proposition 6.3. Suppose that the assumptions of Theorem 2.12 and Assumption 6.2 hold. Fix $z \in \mathcal{D}_{\tau}(\eta_L, \eta_U)$. Recall (6.15). Then for all $i \in [1, N]$, we have that

\[ |P_i(z)| \prec \Psi(z), \quad |K_i(z)| \prec \Psi(z). \]  

(6.24)

Furthermore, we have

\[ \Lambda_d^c(z) \prec \Psi(z), \quad \Lambda_T \prec \Psi(z), \quad \tilde{\Lambda}_d^c \prec \Psi(z), \quad \tilde{\Lambda}_T \prec \Psi(z), \quad \Upsilon \prec \Psi(z). \]  

(6.25)

We now prove Proposition 6.3. As we will see later, the proof of (6.25) follows from (6.24) and (6.23), whereas (6.24) can be proved using recursive estimates.

Proof of Proposition 6.3 We first prove (6.25) assuming that (6.24) holds. We claim that

\[ T_i \prec N^{-\gamma/4}, \]  

(6.26)

where $\gamma > 0$ is the same parameter in (6.23) (also in (2.15)). To see (6.26), using (6.24) and the definition of $K_i$ in (6.15), we have

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

(6.27)

By (6.1), we have that

\[ (\tilde{B}G)_{ii} = \frac{zG_{ii} + 1}{a_i}, \quad G_{ii} = \frac{1}{z^2} \left( a_i(\tilde{B}G)_{ii} - 1 \right). \]  

(6.28)

Based on (6.28), on one hand, by (6.24), we find that

\[ (\tilde{B}G)_{ii} = \frac{1}{a_i - \Omega_B} + O(N^{-\gamma/4}). \]  

(6.29)

On the other hand, using the fact that $\text{tr}(GAB) = zm_H(z) + 1$ and a relation similar to (2.10), by (6.28), we conclude that

\[ \text{tr}(GAB) = \int \frac{x}{x - \Omega_B} \, d\mu_A(x) + O(\Psi(N^{-\gamma/4}) = zm_{\mu_A} \mathbb{E}_{\mu_B}(z) + 1 + O(\Psi(N^{-\gamma/4})), \]  

(6.30)

and

\[ \text{tr}(GA) = \frac{1}{N} \sum_i a_i G_{ii} = \frac{\Omega_B}{z} \sum_i \frac{a_i}{a_i - \Omega_B} + O(\Psi(N^{-\gamma/4}) = \frac{\Omega_B}{z} (zm_{\mu_A} \mathbb{E}_{\mu_B}(z) + 1) + O(\Psi(N^{-\gamma/4}). \]  

(6.31)

Combining (6.27), (6.29), (6.30) and (6.31), by (6.23), we have that

\[ T_i \left( 1 + (zm_{\mu_A} \mathbb{E}_{\mu_B}(z) + 1) \left( \frac{b_i \Omega_B}{z} - 1 \right) \right) = O(\Psi + N^{-\gamma/4}). \]  

(6.32)
Moreover, invoking (2.4) and (2.8), we see that

\[ 1 + \left( zm_{\mu_A, \mu_B}(z) + 1 \right) \left( \frac{b_i \Omega_B}{z} - 1 \right) = \left( zm_{\mu_A, \mu_B}(z) + 1 \right) \left( \frac{b_i \Omega_B}{z} - M_{\mu_A, \mu_B}(z) \right) \]

\[ = \left( zm_{\mu_A, \mu_B}(z) + 1 \right) \Omega_B \frac{(b_i - \Omega_A)}{z}. \]  

(6.33)

By (6.33), a relation similar to (6.11) and (i) of Proposition 3.1, we have proved the claim (6.26) using (6.32).

Armed with (6.26) and (i) of Proposition 3.1, we have proved the claim (6.26) using (6.32).

By (6.24), a relation similar to (2.10) and (i) of Proposition 3.1, we have proved the claim (6.26) using (6.32).

We will give the proof of Lemma 6.4 in the end of this subsection. To conclude the proof of Proposition 6.3, we explain how Lemma 6.4 implies (6.39). We first recall Young's inequality. For any constants \( m \),

\[ \frac{1}{N} \sum_i a_i P_i = Y \frac{1}{N} \sum_i a_i (G_{ii} + T_i) \prec \Psi, \]

(6.34)

where we used the fact that \( \{a_i\} \) are bounded. By (6.34), (6.31), (i) of Proposition 3.1, we have proved that

\[ Y \prec \Psi. \]

Second, using the definition of \( P_i \) in (6.15), the expansion (6.11) and (6.35), we have proved that \( \Lambda_T \prec \Psi. \)

Third, by (6.27) and a discussion similar to (6.26) with the bound \( \Lambda_T \prec \Psi(z) \), it is easy to see that \( \Lambda_T \prec \Psi(z) \).

Finally, the proof of \( \Lambda_T \) and \( \Lambda_T \) follows from an argument similar to (6.11) and the relationship (6.1). This completes the proof of (6.25).

It remains to prove (6.24). Indeed, it suffices to bound moments of \( P_i \) and \( K_i \). More specifically, by Markov inequality, we shall prove for all positive integer \( p \geq 2 \), the followings hold

\[ \mathbb{E} |P_i|^{2p} \prec \Psi^{2p} \quad \text{and} \quad \mathbb{E} |K_i|^{2p} \prec \Psi^{2p}. \]

(6.36)

Denote

\[ X_i^{(p,q)} := P_i^{p} K_i^{q}, \quad \text{and} \quad Y_i^{(p,q)} := \Lambda_T^{p} K_i^{q}. \]

(6.37)

We now state the recursive moment estimates for \( X_i^{(p,q)} \) and \( Y_i^{(p,q)} \).

**Lemma 6.4.** For any fixed integer \( p \geq 2 \) and \( i \in \{1, N\} \), we have

\[ \mathbb{E} X_i^{(p,p)} \leq \mathbb{E} \left[ O_\prec(\Psi) X_i^{(p-1,p)} \right] + \mathbb{E} \left[ O_\prec(\Psi^2) X_i^{(p-2,p-1)} \right], \]

(6.38)

\[ \mathbb{E} Y_i^{(p,p)} \leq \mathbb{E} \left[ O_\prec(\Psi) Y_i^{(p-1,p)} \right] + \mathbb{E} \left[ O_\prec(\Psi^2) Y_i^{(p-2,p-1)} \right]. \]

(6.39)

We will give the proof Lemma 6.4 in the end of this subsection. To conclude the proof of Proposition 6.3, we explain how Lemma 6.4 implies (6.39). We first recall Young’s inequality. For any constants \( \alpha, \beta > 0 \), we have

\[ \alpha \beta \leq \frac{\alpha^m}{m} + \frac{\beta^n}{n}, \quad \frac{1}{m} + \frac{1}{n} = 1, \quad m, n > 1 \text{ are real numbers.} \]

For \( k = 1, 2 \), any arbitrary small constant \( \epsilon > 0 \) and any random variable \( \mathcal{R} = O_\prec(\Psi^k) \) satisfying \( \mathbb{E} [\mathcal{R}]^{q} \prec \Psi^{qk} \), we have that

\[ \mathbb{E} |\mathcal{R} P_i^{2p-k}| \leq \mathbb{E} \left[ N^\epsilon \mathcal{R} |N^{-2p-\epsilon} P_i^{2p-k}\right] \leq \frac{k N^{2p}}{2p} \mathbb{E} \left[ |\mathcal{R}|^{2p} \right] + \frac{(2p - k)N^{-2p}}{2p} \mathbb{E} \left[ |P_i|^{2p-k} \right] + \frac{2p}{2p} \Psi^{2p} + \left( \frac{2p}{2p} \right) N^{-2p} \mathbb{E} \left[ |P_i|^{2p} \right], \]

(6.40)

where in the first inequality we used Young’s inequality with \( m = 2p/k \) and \( n = 2p/(2p - k) \) and in the second equality we used \( \mathbb{E} [\mathcal{R}]^{q} \prec \Psi^{qk} \). Together with (6.38), it yields that

\[ \mathbb{E} |P_i|^{2p} \leq \frac{3}{2p} N^{(2p+1)\epsilon} \Psi^{2p} + \frac{3(2p - 1)}{2p} N^{-\frac{2p}{(2p+1)\epsilon}} \mathbb{E} [\mathcal{R}]^{q} \cdot \frac{2p}{2p} \Psi^{2p} + \left( \frac{2p}{2p} \right) N^{-2p} \mathbb{E} \left[ |P_i|^{2p} \right]. \]
Since $\epsilon > 0$ is arbitrarily small, we can conclude the first part of (6.36). The second part can be proved similarly and we omit the details here.

The rest of the section is devoted to the proof of Lemma 6.4. Throughout the proof, we will need some large deviation estimates as our technical inputs. These estimates and their proof can be found in Lemmas 13.1, 13.2 and 13.3 of Appendix B.

**Proof of Lemma 6.4.** We start with the proof of (6.38). Since $h_{\ell i} e_i^* \tilde{B}^{(i)} Ge_i = b_i h_{\ell i} G_{i i}$, we can rewrite (6.41) as

$$\tilde{B} G = -S_i + \ell_1^2 (b_i h_{\ell i} + h_i^* \tilde{B}^{(i)} h_i)(G_{i i} + T_i) = -S_i + G_{i i} + T_i + e_{i 1},$$

where we denoted

$$e_{i 1} := (\ell_1^2 - 1) b_i h_{\ell i} G_{i i} + (\ell_1^2 h_i^* \tilde{B}^{(i)} h_i - 1)(G_{i i} + T_i) + \ell_1^2 b_i h_{i} T_i.$$  \hfill (6.42)

By Lemma 13.1 and recall $\tilde{g} \sim \mathcal{N}(0, N^{-1} I_N)$, we see that

$$h_{i i} = \|\tilde{g}_i\|^{-1} |e_i^* \tilde{g}_i| < N^{-1/2}.$$  \hfill (6.43)

Consequently, using the definitions in (6.4), we obtain that

$$\ell_1^2 = \frac{2}{\|e_i + h_i\|^2} = \frac{1}{1 + e_i^* h_i} = 1 + O_{\prec}(N^{-1/2}).$$  \hfill (6.44)

Moreover, by (6.4), (6.8), (6.6) and Lemma 13.1 we have

$$h_i^* \tilde{B}^{(i)} h_i = h_i^* R_i \tilde{B} R_i h_i = e_i^* \tilde{B} e_i = \frac{1}{\|\tilde{g}_i\|^2} \tilde{g}_i^* B \tilde{g}_i = tr B + O_{\prec}(N^{-1/2}) = 1 + O_{\prec}(N^{-1/2}),$$  \hfill (6.45)

where we recall that $B$ is normalized such that $tr B = 1$. Using the definition (6.42), by (6.43), (6.44), (6.45) and (6.23), we conclude that

$$|e_{i 1}| < N^{-1/2}.$$  \hfill (6.46)

Therefore, by (6.15), (6.41) and (6.46), we have

$$\mathbb{E} \left[ \chi_i^{(p, p)} \right] = \mathbb{E} \left[ (G_{i i} tr(A \tilde{B}G) + tr(GA)(\tilde{S}_i) + (G_{i i} + T_i)(Y - tr(GA))) \chi_i^{(p-1, p)} \right]$$

$$+ \mathbb{E} \left[ e_{i 1} tr (GA) \chi_i^{(p-1, p)} \right].$$  \hfill (6.47)

Next, we control all the terms of the RHS of (6.47). We mainly focus on the term involving $tr(GA)(\tilde{S}_i)$. As we will see later, by exploring the hidden terms using integration by parts, the term involving $tr(GA)(\tilde{S}_i)$ will generate several terms which would cancel the rest of the terms on the RHS of (6.47). Note that

$$\tilde{S}_i = h_i^* \tilde{B}^{(i)} Ge_i = \sum_k h_{i k}^* e_k^* \tilde{B}^{(i)} Ge_i = \sum_k \frac{1}{\|g_k\|^2} \|g_k\|^2 e_k^* \tilde{B}^{(i)} Ge_i,$$  \hfill (6.48)

where we use the shorthand notation

$$\sum_i \frac{1}{\|g_i\|^2} \|g_i\|^2 = \sum_{k=1}^N \sum_{k \neq i}.$$  

Our calculation relies on the following integration by parts formula for complex centered Gaussian variable $g \sim \mathcal{N}(0, \sigma^2)$ (see eq. (5.33) of [4])

$$\int_{\mathbb{C}} \bar{g} f(g, \bar{g}) e^{-\frac{|g|^2}{2}} d^2 g = \sigma^2 \int_{\mathbb{C}} \partial_{\bar{g}} f(g, \bar{g}) e^{-\frac{|g|^2}{2}} d^2 g,$$  \hfill (6.49)
where \( f : \mathbb{C}^2 \to \mathbb{C} \) is a differentiable function. By (6.48) and (6.49), we have
\[
\mathbb{E} \left[ \bar{S}_i \text{tr}(GA)\mathcal{X}^{(p-1,p)} \right] = \sum_k^{(i)} \mathbb{E} \left[ \frac{1}{\|g_i\|} e_k^{(i)} B^{(i)} G e_i \text{tr}(GA)\mathcal{X}^{(p-1,p)} \right]
\]
\[
= \frac{1}{N} \sum_k^{(i)} \mathbb{E} \left[ \frac{\partial \|g_i\|^{-1}}{\partial g_{ik}} e_k^{(i)} B^{(i)} G e_i \text{tr}(GA)\mathcal{X}^{(p-1,p)} \right] + \frac{1}{N} \sum_k^{(i)} \mathbb{E} \left[ \frac{1}{\|g_i\|} \frac{\partial (e_k^{(i)} B^{(i)} G e_i)}{\partial g_{ik}} \text{tr}(GA)\mathcal{X}^{(p-1,p)} \right]
\]
\[
+ \frac{1}{N} \sum_k \mathbb{E} \left[ \frac{e_k^{(i)} B^{(i)} G e_i}{\|g_i\|} \text{tr}(GA)\mathcal{X}^{(p-1,p)} \right] + \frac{p-1}{N} \mathbb{E} \left[ \frac{1}{\|g_i\|} e_k^{(i)} B^{(i)} G e_i \frac{\partial P_i}{\partial g_{ik}} \mathcal{X}^{(p-2,p)} \right]
\]
\[
+ \frac{p}{N} \mathbb{E} \left[ \frac{1}{\|g_i\|} e_k^{(i)} B^{(i)} G e_i \text{tr}(GA)\mathcal{X}^{(p-1,p-1)} \right],
\]
where we recall that \( \mathcal{X}^{(p-1,p)} = P^{p-1} P^p \) as defined in (6.37). To control the terms on the RHS of (6.50), it suffices to investigate the coefficients involving the partial derivatives with respect to \( g_{ik} \). More specifically, the second term of the RHS of (6.50) will provide some hidden terms for cancellation. Further, since \( B^{(i)} \) is independent of \( v_i \), we have that
\[
\frac{\partial (e_k^{(i)} B^{(i)} G e_i)}{\partial g_{ik}} = e_k^{(i)} B^{(i)} G e_i.
\]

We start with the analysis of the Household reflection defined in (6.4). Recall \( r_i = \ell_i(e_i + h_i) \) and \( \ell_i \) defined in (6.7). By (A.2), (A.3), (A.4) and (A.5), we have that
\[
\frac{\partial R_i}{\partial g_{ik}} = -\ell_i \|g_i\|^{-1} \bar{h}_{ik} h_{ii}(e_i + h_i)(e_i + h_i)^* - \ell_i^2 \|g_i\|^{-1} (e_k e_i^* - \bar{h}_{ik}(h_i e_i^* + e_i h_i^*) + e_k h_i^* - \bar{h}_{ik} h_i h_i^* - \bar{h}_{ik} h_i^*).
\]

We can further rewrite the above equation as
\[
\frac{\partial R_i}{\partial g_{ik}} = -\ell_i \|g_i\|^{-1} e_k(e_i^* + h_i^*) + \Delta_R(i, k),
\]
where we defined
\[
\Delta_R(i, k) := -\ell_i \|g_i\|^{-1} \bar{h}_{ik} h_{ii}(e_i + h_i)(e_i + h_i)^* + \ell_i^2 \|g_i\|^{-1} \bar{h}_{ik}(h_i e_i^* + e_i h_i^* + 2h_i h_i^*).
\]

By (6.51) and (A.1), we obtain that
\[
\frac{\partial G}{\partial g_{ik}} = \ell_i \|g_i\| G A \left( e_k(e_i^* + h_i^*) B^{(i)} R_i + R_i B^{(i)} e_k(e_i^* + h_i^*) \right) G + \Delta_G(i, k),
\]
where
\[
\Delta_G(i, k) := -GA \left( \Delta_R(i, k) B^{(i)} R_i + R_i B^{(i)} \Delta_R(i, k) \right) G.
\]

We see from (6.53) and (B.2) that
\[
\frac{1}{N} \sum_k^{(i)} e_k^{(i)} B^{(i)} \frac{\partial G}{\partial g_{ik}} e_i = \ell_i \|g_i\| \frac{1}{N} \sum_k^{(i)} e_k^{(i)} B^{(i)} G A \left( e_k(e_i^* + h_i^*) B^{(i)} R_i + R_i B^{(i)} e_k(e_i^* + h_i^*) \right) G e_i + O_\prec(\Pi_2^2)
\]
\[
= \ell_i \|g_i\| \frac{1}{N} \sum_k^{(i)} a_k e_k^{(i)} B^{(i)} G e_k(-h_i^* B - e_i^* B) G e_i + e_k^{(i)} B^{(i)} G R_i B^{(i)} e_k(G_i + h_i^* G e_i) + O_\prec(\Pi_2^2)
\]

42
\[ \frac{\ell_i^2}{\| g \|_2} \frac{1}{N} \sum_{k} a_k(\tilde{B}^{(i)} G)_{kk}(-b_i T_i - (BG)_{ii}) + (e_i^T \tilde{B}^{(i)} GAR_i \tilde{B}^{(i)} e_k)(G_{ii} + T_i) + O_{\prec}(\Pi^2), \quad (6.55) \]

where in the second equality we used (6.54) and in the third equality we used (6.53) and (6.52). Moreover, by (6.23) and the assumption that \{a_i\} and \{b_i\} are bounded, we readily see that

\[ \text{tr}(\tilde{A} \tilde{B}^{(i)} G) - \frac{1}{N} \sum_{k} a_k(\tilde{B}^{(i)} G)_{kk} = \frac{1}{N} a_i b_i G_{ii} < \frac{1}{N}, \quad (6.56) \]

and

\[ \text{tr}(\tilde{B}^{(i)} GAR_i \tilde{B}^{(i)} G) - \frac{1}{N} \sum_{k} e_i^T \tilde{B}^{(i)} GAR_i \tilde{B}^{(i)} e_k = \frac{-b_i}{N} e_i^T G A B h_i = \frac{-b_i}{N} e_i^T (z G + 1) h_i < \frac{1}{N}. \quad (6.57) \]

We claim that we can replace \( \tilde{B}^{(i)} \) by \( \tilde{B} \) in (6.56) and (6.57) without changing the error bound in (6.55). Indeed, by the definition of Householder reflection, we have that

\[ \text{tr}(\tilde{A} \tilde{B} G) - \text{tr}(\tilde{A} \tilde{B}^{(i)} G) = \text{tr}(\tilde{A} \tilde{B} G) - \text{tr}(AR_i \tilde{B} r_i G) \]

\[ = \text{tr}(AR_i \tilde{B}^* G B) + \text{tr}(AR_i \tilde{B}^* r_i G) - \text{tr}(AR_i \tilde{B}^* r_i G) \]

\[ = \frac{1}{N} r_i^* \tilde{B}^* AR_i + \frac{1}{N} r_i^* G A B r_i - \frac{1}{N} r_i^* \tilde{B}^* r_i G A r_i. \quad (6.58) \]

Recall that \( r_i = \ell_i (e_i + h_i) \). We have that

\[ \left| \frac{1}{N} r_i^* \tilde{B}^* G A r_i \right| \lesssim \frac{1}{N} \left( \| \sqrt{A} G^* B e_i \| + \| G^* \tilde{B} h_i \| \right) \lesssim \frac{1}{N} \left( e_i G A G^* B e_i + b_i^2 h_i G G^* h_i \right)^{1/2}, \quad (6.59) \]

where in the second inequality we used the fact that \( \| A \| \) is bounded. Moreover, using (6.22), (6.5) and (6.22), it is easy to see that

\[ e_i^* \tilde{B} G A G^* B e_i = e_i^* \sqrt{A} \tilde{B} \sqrt{A} G G^* B e_i = e_i^* H \tilde{G} G^* B e_i \leq \| H \|_2 \text{Im} \tilde{G}_{ii} \frac{\text{Im} \tilde{G}_{ii}}{a_i}, \quad (6.60) \]

where in the last inequality we used the fact that \( \tilde{G} \) is Hermitian and the Ward identity

\[ \sum_{j=1}^{N} |\tilde{G}_{ij}|^2 = (\tilde{G} \tilde{G}^*)_{ii} = \frac{\text{Im} \tilde{G}_{ii}}{\eta}. \]

Similarly, we can show that

\[ b_i^2 h_i G G^* h_i = b_i^2 e_i G G^* e_i = b_i e_i \tilde{G} B G^* e_i \leq b_i \| B \| \frac{\text{Im} \tilde{G}_{ii}}{\eta}. \quad (6.61) \]

Since \( A, B, H \) are bounded, by (6.59), (6.60) and (6.61), we see that

\[ \left| \frac{1}{N} r_i^* \tilde{B} G A r_i \right| \lesssim \frac{1}{N} \left( \frac{\text{Im} \tilde{G}_{ii}}{\eta} + \frac{\text{Im} \tilde{G}_{ii}}{\eta} \right)^{1/2}. \]

By an analogous discussion, we can control the other two terms of the RHS of (6.58) and obtain that

\[ \left| \frac{1}{N} r_i^* G A B r_i \right| \lesssim \frac{1}{N} \left( \frac{\text{Im} \tilde{G}_{ii}}{\eta} + \frac{\text{Im} \tilde{G}_{ii}}{\eta} \right)^{1/2}, \]

\[ \left| \frac{1}{N} r_i^* \tilde{B}^* r_i G A r_i \right| \lesssim \frac{1}{N} \left( \frac{\text{Im} \tilde{G}_{ii}}{\eta} + \frac{\text{Im} \tilde{G}_{ii}}{\eta} \right)^{1/2}. \]
Further, from the spectral decomposition of \( \tilde{H} \) and \( \tilde{H} \), it is clear that that \( \operatorname{Im} \tilde{G}_{ii}/\eta \geq c \) and \( \operatorname{Im} \tilde{G}_{ii}/\eta \geq c \) for some fixed constant \( c > 0 \). This shows that for some constant \( C > 0 \),
\[
\frac{1}{N} \left( \frac{\operatorname{Im} \tilde{G}_{ii} + \operatorname{Im} \tilde{G}_{ii}}{\eta} \right)^{1/2} \leq \frac{C \operatorname{Im} \tilde{G}_{ii} + \operatorname{Im} \tilde{G}_{ii}}{\eta} = C \Pi^2_t.
\]
Together with (6.55), we arrive at
\[
\operatorname{tr}(A\tilde{B}^{(i)}G) = \operatorname{tr}(A\tilde{B}G) + O_\prec(\Pi^2_t). \tag{6.62}
\]
By a discussion similar to (6.62), we can get
\[
\operatorname{tr}(\tilde{B}^{(i)}G\tilde{A}_{t}, \tilde{B}^{(i)}) = \operatorname{tr}(\tilde{B}G\tilde{A}) + O_\prec(\Pi^2_t). \tag{6.63}
\]
Therefore, by (6.55), (6.62) and (6.63), we conclude that
\[
\frac{1}{N} \sum_k e_k^* \tilde{B}^{(i)} \frac{\partial G}{\partial g_{ik}} e_i = \frac{\ell^2_t}{\|g_i\|} \left( \operatorname{tr}(A\tilde{B}G)(-b_i T_i - (\tilde{B})_{ii}) + \operatorname{tr}(\tilde{B}G\tilde{A})(G_{ii} + T_i) \right) + O_\prec(\Pi^2_t). \tag{6.64}
\]
Note that compared to the expansion (6.47), the coefficient in front of \( \operatorname{tr}(A\tilde{B}G) \) is still different. We need further explore the hidden relation. By a discussion similar to (6.64), we have that
\[
\frac{1}{N} \sum_k e_k^* \tilde{B}^{(i)} \frac{\partial G}{\partial g_{ik}} e_i = \frac{\ell^2_t}{\|g_i\|} \left( \operatorname{tr}(GA)(-b_i T_i - (\tilde{B})_{ii}) + \operatorname{tr}(G\tilde{B}A)(G_{ii} + T_i) \right) + O_\prec(\Pi^2_t). \tag{6.65}
\]
In light of (6.64), (6.65) and (6.47), it suffices to control
\[
\operatorname{tr}(GA) \frac{1}{N} \sum_k e_k^* \tilde{B}^{(i)} \frac{\partial G}{\partial g_{ik}} e_i - \operatorname{tr}(A\tilde{B}G) \frac{1}{N} \sum_k e_k^* \frac{\partial G}{\partial g_{ik}} e_i.
\]
Combining (6.64) and (6.65), we have that
\[
\operatorname{tr}(GA) \frac{1}{N} \sum_k e_k^* \tilde{B}^{(i)} \frac{\partial G}{\partial g_{ik}} e_i - \operatorname{tr}(A\tilde{B}G) \frac{1}{N} \sum_k e_k^* \frac{\partial G}{\partial g_{ik}} e_i
\]
\[
= \frac{\ell^2_t}{\|g_i\|} (G_{ii} + T_i)(\operatorname{tr}(GA) \operatorname{tr}(\tilde{B}G\tilde{A}) - \operatorname{tr}(G\tilde{B}A) \operatorname{tr}(G\tilde{B}A)) + O_\prec(\Pi^2_t)
\]
\[
= \frac{\ell^2_t}{\|g_i\|} (G_{ii} + T_i)(-\Upsilon - \operatorname{tr}(A\tilde{B}G) + \operatorname{tr}(GA)) + O_\prec(\Pi^2_t), \tag{6.66}
\]
where in the second equality we employed the definition of \( \Upsilon \) in (6.16). Denote
\[
e_{i_2} := \left( \frac{\ell^2_t}{\|g_i\|} \right)(-G_{ii} \operatorname{tr}(A\tilde{B}G) - (G_{ii} + T_i)(\Upsilon - \operatorname{tr}(GA)) + \operatorname{tr}(A\tilde{B}G) \left( \|g_i\| T_i - \frac{\ell^2_t}{\|g_i\|} T_i \right). \tag{6.67}
\]
By a discussion similar to (6.46), we can conclude that
\[
|e_{i_2}| < N^{-1/2}. \tag{6.68}
\]
Moreover, by a simple algebraic calculation using (6.66) and (6.67), we find that
\[
\operatorname{tr}(GA) \frac{1}{N} \sum_k e_k^* \tilde{B}^{(i)} \frac{\partial G}{\partial g_{ik}} e_i = \|g_i\| (G_{ii} \operatorname{tr}(A\tilde{B}G) - (G_{ii} + T_i)(\Upsilon - \operatorname{tr}(GA)))
\]
44
we find that

to conclude the proof of (6.38), it suffices to control the coefficients

With (6.69) and (6.50), we come back to the discussion (6.37). More specifically, inserting (6.69) into (6.50) and then (6.47), we have that

\[
E \left[ x_i^{(p,p)} \right] = \mathbb{E} \left[ \left( \frac{1}{N} \sum_k e_i^k \frac{\partial G}{\partial g_{ik}} e_i - \|g_i\| \tilde{T}_i \right) \text{tr}(\tilde{A} \tilde{B} G) x_i^{(p-1,p)} \right] + \frac{1}{N} \sum_k \mathbb{E} \left[ e_i^k \tilde{B}^{(i)} G e_i \text{tr}(GA) \frac{\partial \tilde{G}}{\partial g_{ik}} x_i^{(p-2,p)} \right] + \frac{p-1}{N} \sum_k \mathbb{E} \left[ e_i^k \tilde{B}^{(i)} G e_i \text{tr}(GA) \frac{\partial P_i}{\partial g_{ik}} x_i^{(p-1,p-1)} \right]
\]

We do one more expansion for the first term of the above equation. Recall the definitions in (6.13). Applying the technique of integration by parts, i.e., (6.49), we get that

\[
E \left[ \tilde{T}_i - \frac{1}{N} \sum_k e_i^k \frac{\partial G}{\partial g_{ik}} e_i \right] \text{tr}(\tilde{A} \tilde{B} G) x_i^{(p-1,p)} = \frac{1}{N} \sum_k \mathbb{E} \left[ \frac{\partial \|g_i\|^{-1}}{\partial g_{ik}} G_{ki} \text{tr}(\tilde{A} \tilde{B} G) x_i^{(p-1,p)} \right] + \frac{1}{N} \sum_k \mathbb{E} \left[ \frac{1}{\|g_i\|} G_{ki} \text{tr}(\tilde{A} \tilde{B} G) x_i^{(p-1,p)} \right] + \frac{1}{N} \sum_k \mathbb{E} \left[ \text{tr}(\tilde{A} \tilde{B} G) \frac{\partial P_i}{\partial g_{ik}} x_i^{(p-1,p-1)} \right].
\]

Combining (6.71) and (6.70), we can rewrite

\[
E \left[ x_i^{(p,p)} \right] = E \left[ c_1 x_i^{(p-1,p)} \right] + E \left[ c_2 x_i^{(p-2,p)} \right] + E \left[ c_3 x_i^{(p-1,p-1)} \right],
\]

where the coefficients \( c_k, k = 1, 2, 3 \) are defined as

\[
c_1 := \frac{1}{N} \sum_k \left( \frac{\partial \|g_i\|^{-1}}{\partial g_{ik}} G_{ki} \text{tr}(\tilde{A} \tilde{B} G) + \frac{1}{\|g_i\|} G_{ki} \text{tr}(\tilde{A} \tilde{B} G) + \frac{\partial \|g_i\|^{-1}}{\partial g_{ik}} e_i^k \tilde{B}^{(i)} G e_i \text{tr}(GA) \right)
\]

\[
c_2 := \frac{p-1}{N} \sum_k \left( \frac{1}{\|g_i\|} e_i^k \tilde{B}^{(i)} G e_i \text{tr}(GA) \frac{\partial P_i}{\partial g_{ik}} + \frac{1}{\|g_i\|} G_{ki} \text{tr}(\tilde{A} \tilde{B} G) \frac{\partial P_i}{\partial g_{ik}} \right)
\]

\[
c_3 := \frac{p}{N} \sum_k \left( \frac{1}{\|g_i\|} e_i^k \tilde{B}^{(i)} G e_i \text{tr}(GA) \frac{\partial P_i}{\partial g_{ik}} + \frac{1}{\|g_i\|} G_{ki} \text{tr}(\tilde{A} \tilde{B} G) \frac{\partial P_i}{\partial g_{ik}} \right).
\]

To conclude the proof of (6.38), it suffices to control the coefficients \( c_k, k = 1, 2, 3 \). For \( c_1 \), by Lemma \( 3.3 \) we find that

\[
c_1 \ll N^{-1/2} + \Pi_2^2.
\]
For $c_2$, by \((6.36)\) and Lemma \((B.3)\) we find that
\[
c_2 < \Pi^2, \tag{6.77}
\]
Similarly, we can show that
\[
c_3 < \Pi^2, \tag{6.78}
\]
Invoking the definitions in \((6.10)\), we complete the proof of \((6.38)\) using \((6.76)\), \((6.77)\), \((6.78)\) and \((6.72)\).

Finally, due to similarity, we only briefly discuss the proof of \((6.39)\). Using the definition of $K_i$ in \((6.15)\) and the fact that $T_i - T_i = h_iG_{ii} < N^{-1/2}$, we find that
\[
E \left[ \Omega_i^{(p,p)} \right] = E \left[ (\hat{T}_i + \text{tr}(GA)(b_{i}T_i + (\bar{B}G)_{ii}) - \text{tr}(GABG)(G_{ii} + T_i)) \Omega_i^{(p-1,p)} \right] + E \left[ O_{<}(N^{-1/2}) \Omega_i^{(p-1,p)} \right]
\]
\[
= \sum_k E \left[ \frac{\partial g_{ik}}{\partial g_{ik}} G_{ik} e_{k} e_{i} \Omega_i^{(p-1,p)} \right] + E \left[ (\text{tr}(GA)(b_{i}T_i + (\bar{B}G)_{ii}) - \text{tr}(GABG)(G_{ii} + T_i)) \Omega_i^{(p-1,p)} \right] \tag{6.79}
\]
\[
+ E \left[ O_{<}(N^{-1/2}) \Omega_i^{(p-1,p)} \right],
\]
where in the second equality we used the definition in \((6.13)\). Applying \((6.49)\) to the first term of the RHS of the above equation, we obtain
\[
\sum_k E \left[ \frac{\partial g_{ik}}{\partial g_{ik}} G_{ik} e_{k} e_{i} \Omega_i^{(p-1,p)} \right] = \frac{1}{N} \sum_k E \left[ \frac{\partial g_{ik}}{\partial g_{ik}} G_{ik} e_{k} e_{i} \Omega_i^{(p-1,p)} \right] + \frac{1}{N} \sum_k E \left[ \frac{\partial g_{ik}}{\partial g_{ik}} G_{ik} e_{k} e_{i} \Omega_i^{(p-1,p)} \right]
\]
\[
+ \frac{p - 1}{N} \sum_k E \left[ \frac{\partial k_{ik}}{\partial g_{ik}} G_{ik} e_{k} e_{i} \Omega_i^{(p-2,p)} \right] + \frac{p - 1}{N} \sum_k E \left[ \frac{\partial k_{ik}}{\partial g_{ik}} G_{ik} e_{k} e_{i} \Omega_i^{(p-2,p)} \right]. \tag{6.80}
\]
Inserting \((6.65)\) into \((6.80)\) and then \((6.79)\), by a discussion similar to the cancellation in \((6.70)\) and error controls in \((6.46)\) and \((6.68)\), we conclude that
\[
E \left[ \Omega_i^{(p,p)} \right] = \frac{1}{N} \sum_k E \left[ \frac{\partial g_{ik}}{\partial g_{ik}} G_{ik} e_{k} e_{i} \Omega_i^{(p-1,p)} \right] + \frac{p - 1}{N} \sum_k E \left[ \frac{\partial k_{ik}}{\partial g_{ik}} G_{ik} e_{k} e_{i} \Omega_i^{(p-2,p)} \right]
\]
\[
+ \frac{p}{N} \sum_k E \left[ \frac{\partial k_{ik}}{\partial g_{ik}} G_{ik} e_{k} e_{i} \Omega_i^{(p-1,p-1)} \right] + E \left[ O_{<}(N^{-1/2}) \Omega_i^{(p-1,p)} \right].
\]
Using \((A.7)\), Lemma \((B.3)\) and a discussion similar to \((6.70)\), \((6.77)\) and \((6.78)\), we can finish the proof of \((6.39)\).

Before concluding this section, we briefly discuss the control of the off-diagonal entries. Specifically, we will prove the following proposition. Recall the definitions in \((6.21)\).

**Proposition 6.5.** Fix $z \in D_\eta(\eta_L, \eta_U)$. Suppose that the assumptions of Theorem \((2.12)\) Assumption \((6.2)\) hold. Moreover, we assume that
\[
\Lambda_o := \max_{i \neq j} |G_{ij}| < N^{-\gamma/4}, \quad \Lambda_T := \max_{i \neq j} |T_{ij}| < 1. \tag{6.81}
\]
Then for all $i, j \in \mathbb{N}$, we have
\[
|\tilde{P}_{ij}(z)| \prec \Psi, \quad |K_{ij}(z)| \prec \Psi, \quad \Lambda_o \prec \Psi, \quad \Lambda_T \prec \Psi.
\]

**Proof.** The proof is similar to Proposition \((6.3)\) using the identities \((6.14)\), \((6.20)\) and the estimates in Lemma \((B.4)\). We omit further details here. \(\square\)
7 Fluctuation averaging

7.1 Rough fluctuation averaging

As we have seen from Proposition 6.3, the error bounds are not optimal compared to our final results in Theorem 2.12. More specifically, the obtained control for $\Upsilon$ in (6.35) is too loose. In this subsection, we improve the control for $\Upsilon$ based on the Proposition 6.3, which are better estimates compared to (6.23). More specifically, instead of using (6.46) and (6.68), we will conduct a more careful analysis for $e_i$ and $e_i^*$ defined in (6.42) and (6.67), respectively. The improved estimates will be utilized in Section 7.2 to prove Theorem 2.12 for each fixed spectral parameter $z$.

The main result of this section is Proposition 7.2, which provides the estimates for an extension of $\Upsilon$.

More specifically, we will consider the control of the weighted average of $Q_i$’s, i.e.,

$$X(D) := \frac{1}{N} \sum_i d_i Q_i = \operatorname{tr}(GAB) \operatorname{tr}(GD) - \operatorname{tr}(GA) \operatorname{tr}(\tilde{B}GD),$$

where $D = \text{diag}\{d_1, \cdots, d_N\}$ and $d_i \equiv d_i(H)$ are generic weights which in general are functions of $H$. It is necessary and natural for us to consider such a generalization. For instance, in the decomposition (6.11), we have an extra factor, which is a function of $H$, in front of $Q_i$. We first impose some assumptions regarding the concentration properties on $d_i$, $i = 1, 2, \cdots, N$. It will be seen later that the following assumption will be sufficient for most of our applications. Especially, when all $d_i$, $i = 1, 2, \cdots, N$, are functions irrelevant of $H$, Assumption 7.1 will hold trivially.

Assumption 7.1. Let $X_i = I$ or $\tilde{B}^{(i)}$, and let $d_1, \cdots, d_N$ be functions of $H$ with $\max_i |d_i| < 1$. Assume that for all $i, j \in \{1, N\}$ the following hold:

$$\frac{1}{N} \sum_k \frac{\partial d_j}{\partial g_{ik}} e_k^* X_i G e_i = O_{\prec}(\Psi^2 \Pi_i^2),$$

and the same bound also holds with $d_j$’s replaced by $\overline{d}_j$.

Now we state the main result of this subsection.

Proposition 7.2. Fix $z \in D_{\tau}(\eta_L, \eta_U)$ and suppose that the assumptions of Proposition 6.3 and Assumption 7.1 hold. Let $\Pi(z) \prec \tilde{\Pi}(z)$ for some deterministic positive $\tilde{\Pi}(z)$ with $N^{-1/2} \eta^{-1/4} \prec \tilde{\Pi} \prec \Psi$. Then we have

$$X \prec \Psi \tilde{\Pi}.$$

The proof of Proposition 7.2 is similar to that of (6.24), which follows from a recursive estimate Lemma 6.4 and Young’s inequality.

Proof of Proposition 7.2 Denote

$$X^{(p,q)} := X^p X^q, \quad p, q \in \mathbb{N}.$$

We claim that the following recursive estimates hold for $X$.

Lemma 7.3. For any fixed integer $p \geq 2$, we have that

$$E\left[X^{(p,q)}\right] \leq E\left[O_{\prec}(\tilde{\Pi}) X^{(p-1,q)}\right] + E\left[O_{\prec}(\Psi^2 \tilde{\Pi}^2) X^{(p-2,q)}\right] + E\left[O_{\prec}(\Psi^2 \tilde{\Pi}^2) X^{(p-1,q-1)}\right].$$

By a discussion similar to (6.40), together with Lemma 7.3 and Markov inequality, we can complete the proof.
The rest of this section is devoted to proving Lemma 7.3, where we will again apply the technique of integration by parts, i.e., (6.49). The proof is similar to that of Lemma 6.4, except that we have better inputs as demonstrated in Proposition 6.3. In what follows, we focus on discussing the hidden cancellation terms and will only briefly investigate the error terms.

**Proof of Lemma 7.3.** In the first step, we follow the proof idea of [7, Lemma 6.2] to provide a sufficient condition for (7.3); that is, if \( \hat{\Upsilon}(z) \) is another deterministic control parameter such that \( |\Upsilon(z)| \preceq \hat{\Upsilon}(z) \leq \Psi(z) \), then

\[
E \left[ X^{(p,p)} \right] \leq E \left[ O_\prec (\hat{\Pi}^2 + \Psi^2 \hat{\Upsilon}) X^{(p-1,p)} \right] + E \left[ O_\prec (\Psi^2 \hat{\Pi}^2) X^{(p-2,p)} \right] + E \left[ O_\prec (\Psi^2 \hat{\Pi}^2) (X^{(p-1,p-1)}) \right].
\]  

(7.4)

Similar to the discussion of (6.40), we can employ Young’s and Markov inequalities to obtain that

\[
\left| \sum \frac{1}{N} d_i Q_i \right| \prec \hat{\Pi}^2 + \Psi \hat{\Upsilon} + \Psi \hat{\Pi} \prec \Psi \hat{\Upsilon} + \Psi \hat{\Pi},
\]

(7.5)

where we used the assumption \( \hat{\Pi} \prec \Psi \). Recall (6.17). We now set \( d_i = z \) for all \( i \) in (7.4) to get

\[
|\Upsilon(z)| \prec \Psi \hat{\Upsilon} + \Psi \hat{\Pi} \prec N^{-\gamma/2} \hat{\Upsilon}(z) + \Psi \hat{\Pi},
\]

(7.6)

where we used the fact that \( z \in D_\tau(\eta_L, \eta_U) \) is fixed. Using the RHS of (7.6) as an updated deterministic bound for \( \Upsilon \) instead of \( \hat{\Upsilon} \) and iterating the estimation procedure as in (7.3), we can finally obtain that

\[
|\Upsilon(z)| \prec \Psi \hat{\Pi}.
\]

Consequently, we can set \( \hat{\Upsilon}(z) = \Psi \hat{\Pi} \) in (7.3). This yields that

\[
E \left[ X^{(p,p)} \right] \leq E \left[ O_\prec (\hat{\Pi}^2 + \Psi^2 \hat{\Pi}) X^{(p-1,p)} \right] + E \left[ O_\prec (\Psi^2 \hat{\Pi}^2) (X^{(p-2,p)} + X^{(p-1,p-1)}) \right].
\]

(7.7)

Since the term \( \Psi^2 \hat{\Pi} \) in the first expectation of the RHS of (7.7) can be absorbed into \( \hat{\Pi}^2 \) as we assume that \( N^{-\gamma/2} \eta_{-1/4} \prec \hat{\Pi} \), this recovers (7.3).

It remains to prove (7.4). Recall \( D = \text{diag}\{d_1, \ldots, d_N\} \) and (6.12). We see that

\[
\mathbf{X} = \frac{1}{N} \sum_i d_i Q_i = \frac{1}{N} \sum_i d_i (G_{ii} \text{ tr}(A\hat{B}G) - \text{ tr}(GA)(\hat{B}G)_{ii})
\]

\[
= \frac{1}{N} \sum_i (\hat{B}G)_{ii} \tau_{i1} \text{ tr}(GA),
\]

(7.8)

where we denoted

\[
\tau_{i1} := \frac{a_i \text{ tr}(GD)}{\text{ tr}(GA)} - d_i.
\]

(7.9)

We state two important properties regarding \( \tau_{i1} \). First,

\[
\sum_i G_{ii} \tau_{i1} = \frac{1}{\text{ tr}(GA)} \sum_i G_{ii} a_i \text{ tr}(GD) - \text{ tr}(GD) = 0.
\]

(7.10)

Second, using the identity (c.f. Definition 6.1), we have that

\[
\text{ tr } (GA) = \frac{(1 + z m_H(z)) \Omega_B^\tau(z)}{z}.
\]

Therefore, by Lemma 5.2, Propositions 5.1 and 6.3 it is easy to see that \( \tau_{i1} \prec 1 \).
Inserting (6.41) into (7.8), we obtain that
\[
E \left[ x^{(p,p)} \right] = \frac{1}{N} \sum_{i=1}^{N} E \left[ (\tilde{B}G)_{ii} \tau_{i1} \text{tr}(GA)x^{(p-1,p)} \right]
\]
\[
= -\frac{1}{N} \sum_{i=1}^{N} E \left[ \tilde{S}_i \tau_{i1} \text{tr}(GA)x^{(p-1,p)} \right] + \frac{1}{N} \sum_{i=1}^{N} E \left[ (G_{ii} + T_{ii}) \tau_{i1} \text{tr}(GA)x^{(p-1,p)} \right] 
\]
\[
+ \frac{1}{N} \sum_{i=1}^{N} E \left[ e_{i1} \tau_{i1} \text{tr}(GA)x^{(p-1,p)} \right],
\]  
(7.11)
where \(e_{i1}\) is defined in (6.42). It suffices to estimate all the terms of the RHS of (7.11) and (7.12). Due to (7.10), for part of the second term of the RHS of (7.11), we have that
\[
\frac{1}{N} \sum_{i=1}^{N} E G_{ii} \tau_{i1} \text{tr}(GA)x^{(p-1,p)} = 0.
\]  
(7.13)
For the remaining terms, we apply (6.49) to estimate them. We start with the first term of the RHS of (7.11). Recall (6.43). We have that
\[
E \left[ \tilde{S}_i \tau_{i1} \text{tr}(GA)x^{(p-1,p)} \right] = \frac{1}{N} \sum_{k=1}^{(i)} E \left[ \frac{1}{\|g_i\|} \partial (e_k^i \tilde{B}^{(i)} G e_i) \right] \tau_{i1} \text{tr}(GA)x^{(p-1,p)} 
\]  
(7.14)
\[
+ \frac{1}{N} \sum_{k=1}^{(i)} E \left[ \frac{\partial \|g_i\|^{-1}}{\partial g_{ik}} e_k^i \tilde{B}^{(i)} G e_i \tau_{i1} \text{tr}(GA)x^{(p-1,p)} \right] + \frac{1}{N} \sum_{k=1}^{(i)} E \left[ \frac{\|g_i\|}{\partial g_{ik}} \frac{\partial (\tau_{i1} \text{tr}(GA) x^{(p-1,p)})}{\partial g_{ik}} \right].
\]
For the first term of the RHS of (7.14), by (6.69), we find that
\[
\frac{1}{N^2} \sum_{i=1}^{N} \sum_{k=1}^{(i)} E \left[ \frac{1}{\|g_i\|} \partial (e_k^i \tilde{B}^{(i)} G e_i) \tau_{i1} \text{tr}(GA)x^{(p-1,p)} \right] 
\]
\[
= \frac{1}{N} \sum_{i=1}^{N} E \left[ (-G_{ii} \text{tr}(A \tilde{B} G) - (G_{ii} + T_{ii}) (\Upsilon(z) - \text{tr}(GA))) \tau_{i1} x^{(p-1,p)} \right] 
\]  
(7.15)
\[
+ \frac{1}{N} \sum_{i=1}^{N} E \left[ \text{tr}(A \tilde{B} G) \left( \frac{1}{N^2} \sum_{i=1}^{N} \frac{1}{\|g_i\|} e_k^i \frac{\partial G}{\partial g_{ik}} e_i - \tilde{T}_i \right) \right] \tau_{i1} x^{(p-1,p)} 
\]
\[
+ \frac{1}{N} \sum_{i=1}^{N} E \left[ \frac{\tau_{i1}}{\|g_i\|} O_{\tilde{z}}(\Pi_i^2) x^{(p-1,p)} \right],
\]
where \(e_{i2}\) is defined in (6.67).
Recall (6.22). Using (7.10), \(A_T \prec \Psi(z)\) in Proposition 6.3 and the assumption \(|\Upsilon(z)| \prec \tilde{\Upsilon}(z)\), we conclude that the first term on the RHS of (7.15) can be simplified as
\[
\frac{1}{N} \sum_{i=1}^{N} (-G_{ii} \text{tr}(A \tilde{B} G) - (G_{ii} + T_{ii}) (\Upsilon(z) - \text{tr}(GA))) \tau_{i1} = \frac{1}{N} \sum_{i=1}^{N} T_{ii} \tau_{i1} \text{tr}(GA) + O_{\tilde{z}}(\tilde{\Upsilon}(z) \Psi(z)).
\]  
(7.16)
For the last term of the RHS of (7.15), since \(\tau_{i1} < 1\) and \(\|g_i\| = 1 + O_{\tilde{z}}(N^{-1/2})\), together with the fact
\[
\frac{1}{N} \sum_{i=1}^{N} \Pi_i^2 \prec \Pi^2,
\]
49
Moreover, using (B.1), we obtain that
\[
\frac{1}{N} \sum_i E \left[ \frac{\tau_{i1}}{\|g_i\|} O_\omega (\Pi_i^2) \mathbf{x}^{(p-1,p)} \right] < \Pi^2 < \tilde{\Pi} \Psi.
\]

For the second term of the RHS of (7.15), by a discussion similar to (6.71), we have that
\[
\mathbb{E} \left[ \left( \tilde{T}_{i1} - \frac{1}{N} \sum_k e_i^k \frac{\partial G}{\partial g_{ik}} e_i g_{ik} \right) \tau_{i1} \text{tr}(A\tilde{B}G) \mathbf{x}_i^{(p-1,p)} \right] = \frac{1}{N} \sum_k \mathbb{E} \left[ \frac{\partial \|g_i\|^{-1}}{\partial g_{ik}} G_{ki} \tau_{i1} \text{tr}(A\tilde{B}G) \mathbf{x}_i^{(p-1,p)} \right] + \frac{1}{N} \sum_k \mathbb{E} \left[ G_{ki} \tau_{i1} \text{tr}(A\tilde{B}G) \frac{\partial \mathbf{x}_i}{\partial g_{ik}} \tau_{i1} \mathbf{x}_i^{(p-1,p-1)} \right]
\]
\[
= \frac{p-1}{N} \sum_k \mathbb{E} \left[ G_{ki} \tau_{i1} \text{tr}(A\tilde{B}G) \frac{\partial \mathbf{x}_i}{\partial g_{ik}} \tau_{i1} \mathbf{x}_i^{(p-2,p-1)} \right] + \frac{p}{N} \sum_k \mathbb{E} \left[ G_{ki} \tau_{i1} \text{tr}(A\tilde{B}G) \frac{\partial \mathbf{x}_i}{\partial g_{ik}} \tau_{i1} \mathbf{x}_i^{(p-1,p-1)} \right].
\]

At this point, we deviate a little bit from the current calculation and revisit (7.12). On one hand, we emphasize that, combining (7.10) and (7.13), we could cancel the second term of the RHS of (7.11); on the other hand, for the term of the RHS of (7.12), by the definition of $e_i$, (6.43), (6.44), (6.45) and Proposition 6.3 we have that
\[
e_i = \left( (l_i^2 - 1) + (h_i^* \tilde{B}^{(i)} h_i - 1) \right) G_{ii} + O_\omega (N^{-1/2} \Psi)
\]
\[
= -h_{ii} G_{ii} + \frac{\Omega_B h_{ii} \tilde{B}^{(i)} h_i - 1}{a_i - \Omega_B^c} + O_\omega (N^{-1/2} \Psi).
\]

Recall (6.7). By (B.1), we see that
\[
h_i^* \tilde{B}^{(i)} h_i - h_i^* \tilde{B}^{(i)} h_i = O_\omega (N^{-1}).
\]

Moreover, using (15.1), we obtain that
\[
\hat{h}_i^* \tilde{B}^{(i)} h_i - 1 = h_i^* (\tilde{B}^{(i)} - I) h_i + O_\omega (1/N).
\]

By (7.18), (7.19), (7.20) and the assumption that $N^{-1/2} q^{-1/4} \ll \tilde{\Pi}$, we conclude that the last term in (7.12) can be written as
\[
\frac{1}{N} \sum_i E \left[ \tau_{i1} \text{tr}(GA) \mathbf{x}^{(p-1,p)} \right] = -E \left[ \frac{1}{N} \sum_i f_{ii} G_{ii} \tau_{i1} \text{tr}(GA) \mathbf{x}^{(p-1,p)} \right] + \frac{1}{z} E \left[ \frac{1}{N} \sum_i (h_i^* (\tilde{B}^{(i)} - I) h_i) \tau_{i2} \mathbf{x}^{(p-1,p)} \right] + E \left[ O_\omega (\tilde{\Pi}^2) \mathbf{x}^{(p-1,p)} \right],
\]

where $\tau_{i2}$ is defined as
\[
\tau_{i2} := \frac{\Omega_B^c (a_i \text{tr}(GD) - d_i \text{tr}(GA))}{a_i - \Omega_B^c}.
\]

We point out that the first term of the RHS of (7.21) will be kept for future cancellation (see the first term of the RHS of (7.24)). For the second term of the RHS of (7.21), by (6.49), we see that
\[
\frac{1}{N} \sum_i E \left[ h_i^* (\tilde{B}^{(i)} - I) h_i \tau_{i2} \mathbf{x}^{(p-1,p)} \right] = \frac{1}{N^2} \sum_i \sum_k \mathbb{E} \left[ g_{ik} \frac{1}{\|g_i\|^2} e_i^k (\tilde{B}^{(i)} - I) g_{ik} \tau_{i2} \mathbf{x}^{(p-1,p)} \right].
\]
\[
= \frac{1}{N^2} \sum_{i} \sum_{k} \left[ \frac{\partial||g_i||^{-2}}{\partial g_{ik}} e_k^*(\tilde{B}^{(i)} - I) \tilde{g}_i \tau_2 \mathcal{X}^{(p-1,p)} \right] + \frac{1}{N^2} \sum_{i} \sum_{k} \left[ \frac{\partial||g_i||^{-2}}{\partial g_{ik}} e_k^* \frac{e_k}{||g_i||^2} \tau_2 \mathcal{X}^{(p-1,p)} \right] + \frac{1}{N^2} \sum_{i} \sum_{k} \left[ \frac{\partial||g_i||^{-2}}{\partial g_{ik}} e_k^* \frac{e_k}{||g_i||^2} \tau_2 \mathcal{X}^{(p-2,p)} \right] + \frac{p}{N^2} \sum_{i} \sum_{k} \left[ e_k^* \frac{e_k}{||g_i||^2} \tau_2 \mathcal{X}^{(p-1,p)} \right].
\]

For the first term of the RHS of (7.22), using the facts \(\frac{\partial||g_i||^{-2}}{\partial g_{ik}} = -\frac{1}{||g_i||^4} \tilde{g}_{ik}\) and \(\tau_2 \prec 1\), we can get
\[
\frac{1}{N^2} \sum_{i} \sum_{k} \left[ \frac{\partial||g_i||^{-2}}{\partial g_{ik}} e_k^*(\tilde{B}^{(i)} - I) \tilde{g}_i \tau_2 \mathcal{X}^{(p-1,p)} \right] = -\frac{1}{N^2} \sum_{i} \left[ \frac{1}{||g_i||^4} \tilde{g}_i^* (\tilde{B}^{(i)} - I) \tilde{g}_i \tau_2 \mathcal{X}^{(p-1,p)} \right] = -\mathbb{E} \left[ O_\prec (N^{-1}) \mathcal{X}^{(p-1,p)} \right].
\]

For the second term of the RHS of (7.22), using that \(\text{tr } B = \text{tr } \tilde{B}^{(i)} = 1\), we conclude
\[
\frac{1}{N^2} \sum_{i} \sum_{k} \left[ \frac{1}{||g_i||^2} e_k^* (\tilde{B}^{(i)} - I) e_k \tau_2 \mathcal{X}^{(p-1,p)} \right] = \frac{1}{N^2} \sum_{i} \left[ \frac{1}{||g_i||^2} \tau_2 \mathcal{X}^{(p-1,p)} \right] = \mathbb{E} \left[ O_\prec (1/N) \mathcal{X}^{(p-1,p)} \right].
\]

Now we return to (7.15) and investigate the remaining term, i.e., the third term. Recall (6.67). By a discussion similar to (6.69), we find that
\[
e_{22} = (||g_i||^2 - 1 + h_{ii}) G_{ii} (\text{tr}(A\tilde{B}G) - \text{tr}(GA)) - h_{ii} G_{ii} \text{tr}(A\tilde{B}G) + O_\prec (N^{-1/2}\Psi)
= (||g_i||^2 - 1) G_{ii} \text{tr}(A\tilde{B}G) - h_{ii} G_{ii} \text{tr}(GA) + O_\prec (N^{-1/2}\Psi)
= (||\tilde{g}_i||^2 - 1) \frac{\Omega_B^2 (\text{tr}(A\tilde{B}G))}{a_i - \Omega_B} - h_{ii} G_{ii} \text{tr}(GA) + O_\prec (N^{-1/2}\Psi). \tag{7.23}
\]

Using the definitions in (6.7), by (3.31), (7.23) and a discussion similar to (7.22), we find that the third term of the RHS of (7.21) can be written as
\[
\frac{1}{N} \sum_{i} \mathbb{E} \left[ \left( \frac{6 \tau_{13}}{||g_i||^2} \right) \mathcal{X}^{(p-1,p)} \right] = \frac{1}{N} \sum_{i} \mathbb{E} \left[ \left( ||\tilde{g}_i||^2 - 1 \right) \tau_{13} \mathcal{X}^{(p-1,p)} \right] \tag{7.24}
\]
\[
- \frac{1}{N} \sum_{i} \mathbb{E} \left[ h_{ii} G_{ii} \tau_{13} \text{tr}(GA) \mathcal{X}^{(p-1,p)} \right] + \mathbb{E} \left[ O_\prec (N^{-1/2}\Psi) \mathcal{X}^{(p-1,p)} \right],
\]
\text{where we denoted}
\[
\tau_{13} := \tau_{2} \text{tr}(A(\tilde{B} - I)G).
\]

We emphasize that the first term of the RHS of (7.24) is canceled out with the first term of the RHS of (7.21). Till now, all the leading terms in the expansion (7.12) have been canceled out. Regarding the second term of (7.24), by (6.49) and the definitions in (6.7), we conclude that
\[
\frac{1}{N} \sum_{i} \mathbb{E} \left[ \left( ||\tilde{g}_i||^2 - 1 \right) \tau_{13} \mathcal{X}^{(p-1,p)} \right] = \frac{1}{N^2} \sum_{i} \sum_{k} \left[ e_k^* e_k - 1 + \frac{1}{N - 1} \right] \tau_{13} \mathcal{X}^{(p-1,p)}
\]
\begin{align}
&+ \frac{1}{N^2} \sum_{i} \sum_{k} \sum_{(i)} E \left[ e_k g_i \frac{\partial \tau_{i3}}{\partial g_{ik}} \mathcal{X}^{(p-1,p)} \right] + \frac{p-1}{N^2} \sum_{i} \sum_{k} \sum_{(i)} E \left[ e_k g_i \frac{\partial \mathcal{X}}{\partial g_{ik}} \mathcal{X}^{(p-2,p)} \right] \\
&+ \frac{p}{N^2} \sum_{i} \sum_{k} \sum_{(i)} E \left[ e_k^* \hat{g}_i \frac{\partial \mathcal{X}}{\partial g_{ik}} \mathcal{X}^{(p-1,p-1)} \right].
\end{align}
(7.25)

It is easy to see that the first term of the RHS of (7.25) can be bounded by \( E \left[ O(N^{-1}) \mathcal{X}^{(p-1,p)} \right] \).

Armed with the discussions between (7.13) and (7.25), by a discussion similar to (6.72), we can write (7.25) as follow
\begin{align}
E \left[ \mathcal{X}^{(p,p)} \right] &= E \left[ \mathcal{D}_1 \mathcal{X}^{(p-1,p)} \right] + E \left[ \mathcal{D}_2 \mathcal{X}^{(p-2,p)} \right] + E \left[ \mathcal{D}_3 \mathcal{X}^{(p-1,p-1)} \right],
\end{align}
(7.26)

where \( \mathcal{D}_k, k = 1, 2, 3 \), are defined as
\begin{align*}
\mathcal{D}_1 &= - \sum_{i} \left( \frac{1}{N} \sum_{k} \frac{\partial \| g_i \|^{-1}}{\partial g_{ik}} e_k^* \tilde{B}^{(i)} G e_i \tau_{i1} \text{tr}(GA) - \frac{1}{N} \sum_{k} \frac{e_k^* \tilde{B}^{(i)} G e_i \partial (\tau_{i1} \text{tr}(GA))}{\| g_i \|} \right) \\
&+ \sum_{i} \left( \frac{1}{N} \sum_{k} \frac{\partial \| g_i \|^{-1}}{\partial g_{ik}} G_k i \tau_{i1} \text{tr}(\bar{A} \bar{B} G) \tau_{i1} + \frac{1}{N} \sum_{k} \frac{G_{ki} \partial (\tau_{i1} \text{tr}(\bar{A} \bar{B} G))}{\| g_i \|} + \frac{1}{Nz^2} \sum_{k} \frac{e_k^* (\tilde{B}^{(i)} - I) \hat{g}_i \partial \tau_{i2}}{\| g_i \|^2 \partial g_{ik}} \right) \\
&- \frac{1}{N^2} \sum_{i} \sum_{k} \sum_{(i)} e_k^* g_i \frac{\partial \tau_{i3}}{\partial g_{ik}} + O_N \left( N^{-1} + \bar{\Psi} \right),
\end{align*}
\begin{align*}
\mathcal{D}_2 &= \frac{p}{N} \sum_{i} \left( - \sum_{k} \frac{e_k^* \tilde{B}^{(i)} G e_i \tau_{i1} \text{tr}(GA)}{\| g_i \|} \frac{\partial \mathcal{X}}{\partial g_{ik}} + \sum_{k} \frac{G_{ki} \tau_{i1} \text{tr}(\bar{A} \bar{B} G)}{\| g_i \|} \frac{\partial \mathcal{X}}{\partial g_{ik}} \right) \\
&+ \frac{p-1}{N^2} \sum_{i} \left( \frac{1}{z \sum_{k}} \sum_{(i)} \frac{e_k^* (\tilde{B}^{(i)} - I) \hat{g}_i \tau_{i2}}{\| g_i \|^2} \frac{\partial \mathcal{X}}{\partial g_{ik}} - \sum_{k} \frac{e_k^* \hat{g}_i \tau_{i3}}{\partial g_{ik}} \right),
\end{align*}
\begin{align*}
\mathcal{D}_3 &= \frac{p}{N} \sum_{i} \left( - \sum_{k} \frac{e_k^* \tilde{B}^{(i)} G e_i \tau_{i1} \text{tr}(GA)}{\| g_i \|} \frac{\partial \mathcal{X}}{\partial g_{ik}} + \sum_{k} \frac{G_{ki} \tau_{i1} \text{tr}(\bar{A} \bar{B} G)}{\| g_i \|} \frac{\partial \mathcal{X}}{\partial g_{ik}} \right) \\
&+ \frac{p}{N^2} \sum_{i} \left( \frac{1}{z \sum_{k}} \sum_{(i)} \frac{e_k^* (\tilde{B}^{(i)} - I) \hat{g}_i \tau_{i2}}{\| g_i \|^2} \frac{\partial \mathcal{X}}{\partial g_{ik}} - \sum_{k} \frac{e_k^* \hat{g}_i \tau_{i3}}{\partial g_{ik}} \right).
\end{align*}

Finally, we need to control \( \mathcal{D}_k, k = 1, 2, 3 \), using a discussion similar to (6.73), (6.74) and (6.75) utilizing Lemmas 3.2 and 3.3. Due to similarity, we only sketch the proof and omit the details. Specifically, using the linear property of \( \mathcal{X} \) (c.f. (6.11)) and chain rules, we find that all the terms in \( \mathcal{D}_k, k = 1, 2, 3 \), are linear combinations of the form
\[ \frac{1}{N^2} \sum_{i} \sum_{k} \sum_{(i)} c_i \Gamma_{i,k}, \]
where \( c_i \)'s are some generic weights and \( \Gamma_{i,k} \) can be one of the following expressions
\begin{align}
&e_k^* X_i G e_i \text{tr}(X_D X_A \frac{\partial G}{\partial g_{ik}}) \mathcal{X}^3, \quad e_k^* \hat{g}_i \text{tr}(X_D X_A \frac{\partial G}{\partial g_{ik}}) \mathcal{X}^2, \quad \frac{\partial \| g_i \|^{-1}}{\partial g_{ik}} e_k^* X_i G e_i \mathcal{X}^{(p-1,p)}, \quad (7.27)
\end{align}

52
where $X_i$ is either $\tilde{B}^{(i)}$ or $I$, $X_A$ is $A$, $I$ or $A^{-1}$, $X_D$ is either $D$ or $I$, and $J$ is $(p-1, p)$, $(p-2, p)$, or $(p-1, p-1)$. All the terms in (7.27) can be tackled using Lemma 6.3 whereas the terms in (7.28) can be bounded using Proposition 6.3 the assumption (7.2), Lemmas B.2 and B.3 with the following linear property

$$\text{tr}(\frac{\partial D}{\partial y_{ik}} X_A G) = \frac{1}{N} \sum \frac{\partial d_i}{\partial y_{ik}} (X_A G)_{ij}.$$ 

Based on the above discussion, we can show that

$$\mathcal{D}_1 \prec \tilde{\Pi}^2 + \Psi \tilde{\Pi}, \quad \mathcal{D}_2 \prec \Psi^2 \tilde{\Pi}^2, \quad \mathcal{D}_3 \prec \Psi^2 \tilde{\Pi}^2.$$ 

This completes the proof of (7.31). \hfill \square

### 7.2 Optimal fluctuation averaging

In this section, we will establish an estimate for the key quantities regarding the stability of system which masters the subordination functions (c.f. (7.29)). Such an estimation relies on strengthening the estimates obtained in Proposition 7.2 for certain explicit choices of functions $d_i, i = 1, 2, \cdots, N$; see (7.31) and (7.32) for details. These results will be a base for the proof of Theorem 2.1. Moreover, as one important byproduct, we can show the closeness between the subordination functions and their approximations in Definition 6.1.

Denote

$$\Lambda_A(z) := \Omega_A(z) - \Omega^e_A(z), \quad \Lambda_B(z) := \Omega_B(z) - \Omega^e_B(z), \quad \Lambda(z) := |\Lambda_A(z)| + |\Lambda_B(z)|.$$ 

As we have seen in Section 3, especially Lemma 3.14, the control of $\Lambda$ are expected to reduced to studying a system similar to (3.1). Recall that $\Phi_{AB} \equiv (\Phi_A, \Phi_B) \in \mathbb{C}^3$ and the subordination functions $\Omega_A$ and $\Omega_B$ are governed by the following system

$$\Phi_{AB}(\Omega_A(z), \Omega_B(z), (z)) = 0,$$ 

where $\Phi_A(z)$ and $\Phi_B(z)$ are defined as follows

$$\Phi_A(\omega_1, \omega_2, z) := \frac{M_{\mu_A}(\omega_2)}{\omega_2} \cdot \frac{\omega_1}{z}, \quad \text{and} \quad \Phi_B(\omega_1, \omega_2, z) := \frac{M_{\mu_B}(\omega_1)}{\omega_1} \cdot \frac{\omega_2}{z}. \quad (7.30)$$

We further introduce the following shorthand notations

$$\Phi^e_A \equiv \Phi^e_A(z) := \Phi_A(\Omega^e_A(z), \Omega^e_B(z), z) \quad \text{and} \quad \Phi^e_B \equiv \Phi^e_B(z) := \Phi_B(\Omega^e_A(z), \Omega^e_B(z), z). \quad (7.31)$$

Whenever there is no ambiguity, we will omit the dependence on $z$ and we will often consider $(\Phi_A, \Phi_B)$ as a function of the first two variables $(\omega_1, \omega_2)$. Recall that $S_{AB}, T_A$ and $T_B$ are defined analogously as in (8.3) and (8.4) by replacing the pair $(\alpha, \beta)$ with $(A, B)$. We first introduce an estimate regarding the linear combination of $S_{AB}, T_A, T_B$ and $\Lambda$. It serves as a fundamental input for the continuity argument in Section 8. Before stating the result, we observe that, under Assumption 6.2 by (6.11) and Proposition 6.3

$$\Lambda \sim N^{-\gamma/4}. \quad (7.32)$$

**Proposition 7.4.** Fix $z \in D_\varepsilon(\eta L, \eta_\nu)$. Suppose the assumptions of Proposition 6.3 hold. Let $\tilde{\Lambda}(z)$ be a deterministic positive function such that $\Lambda(z) \ll \tilde{\Lambda}(z) \ll N^{-\gamma/4}$. Then we have for $\tau = A, B,$

$$\left| \frac{S_{AB}(z)}{z} \Lambda_\varepsilon + T_A\Lambda_\varepsilon^2 + O(|\Lambda_\varepsilon|^3) \right| \prec \tilde{\mathcal{U}}. \quad (7.33)$$

where $\mathcal{U}$ is defined as

$$\mathcal{U} \equiv \mathcal{U}(z) := \Psi^2 \left( \sqrt{\text{Im} m_\mu B_{\mu B}(z + \tilde{\Lambda}(z))(|S_{AB}(z)| + \tilde{\Lambda}(z)) + \Psi^2} \right). \quad (7.34)$$

53
The proof of Proposition 7.4 relies on the estimate for a special linear combinations of $Q_i$’s and their analogues. Recall (2.4). Denote

$$Z_1 := \Phi_A^c + zL_{\mu_A}^\prime(\Omega_B)\Phi_B^c, \quad Z_2 := \Phi_A + zL_{\mu_B}^\prime(\Omega_A)\Phi_A^c, \quad Z_1^{(p,q)} := Z_1^p Z_2^q,$$

(7.35)

We collect the recursive moment estimates for $Z_1$ and $Z_2$ in the following lemma. Its proof will be provided after we finish proving Proposition 7.4.

**Lemma 7.5.** Fix $z \in D_r(\eta_L, \eta_U)$. Suppose the assumptions of Proposition 7.4 hold. Then we have

$$\mathbb{E} \left[ Z_1^{(p,p)} \right] = \mathbb{E} \left[ O_{\infty}(\mathbb{U}) Z_1^{(p-1,p)} \right] + \mathbb{E} \left[ O_{\infty}(\mathbb{U}^2) Z_1^{(p-2,p)} \right] + \mathbb{E} \left[ O_{\infty}(\mathbb{U}) Z_1^{(p-1,p-1)} \right],$$

(7.36)

where $\mathbb{U}$ is defined in (7.34). Similar results hold for $Z_2$.

Armed with Lemma 7.5, we are ready to prove Proposition 7.4.

**Proof of Proposition 7.4.** Due to similarly, we will focus on the discussion when $\iota = A$. First of all, we provide some identities to connect the left-hand side of (7.35) and $Z_1$ and $Z_2$. Observing (7.35), by a discussion similar to 3.5, we expand $(\Phi_A, \Phi_B)$ around $(\Omega_A, \Omega_B)$ and obtain

$$\Phi_A^c = L_{\mu_A}^\prime(\Omega_B)\Lambda_B + \frac{1}{2} L_{\mu_A}''(\Omega_B)\Lambda_B^2 - \frac{\Lambda_A}{z} + O(\Lambda_A^3),$$

(7.37)

$$\Phi_B^c = L_{\mu_B}^\prime(\Omega_A)\Lambda_A + \frac{1}{2} L_{\mu_B}''(\Omega_A)\Lambda_A^2 - \frac{\Lambda_B}{z} + O(\Lambda_B^3).$$

(7.38)

Combine (7.35) and (7.37), by a simple algebraic calculation (i.e., inserting (7.35) in terms of $\Lambda_B$ into (7.37)), we can conclude that

$$\Phi_A^c + zL_{\mu_A}^\prime(\Omega_B)\Phi_B^c = \left( zL_{\mu_B}^\prime(\Omega_A)L_{\mu_A}^\prime(\Omega_B) - \frac{1}{z} \right)\Lambda_A + \frac{1}{2} \left( zL_{\mu_B}''(\Omega_A)L_{\mu_A}^\prime(\Omega_B) + (zL_{\mu_B}^\prime(\Omega_A))^2 L_{\mu_A}''(\Omega_B) \right)\Lambda_A^2$$

$$+ O(|\Phi_B|^2) + O(|\Phi_B\Lambda_A|) + O(|\Lambda_A^3|)$$

$$= \frac{S_{AB}}{z} \Lambda_A + T_A \Lambda_A^2 + O(|\Lambda_A|^3) = 3_1 + O_{\infty}(|\Phi_B|^2 + |\Phi_B\Lambda_A|) + O(|\Lambda_A|^3).$$

(7.39)

This implies that

$$\frac{S_{AB}}{z} \Lambda_A + T_A \Lambda_A^2 + O(|\Lambda_A|^3) = 3_1 + O_{\infty}(|\Phi_B|^2 + |\Phi_B\Lambda_A|).$$

(7.40)

By a discussion similar to (6.40), using Young’s and Markov’s inequalities, together with Lemma 7.5, we have that

$$3_1 \sim \mathbb{U}, \quad 3_2 \sim \mathbb{U}.$$

(7.41)

In what follows, we will apply Proposition 4.2 to prove that

$$\Phi_i^c \sim \Psi \mathbb{U}^{1/2}, \quad i = A, B.$$

(7.42)

Recall (6.12) and denote

$$Q_i := \text{tr}(G\tilde{A}B)G_{ii} - \text{tr}(B\tilde{G})G_{ii}.$$

We next write $\Phi_A^c$ and $\Phi_B^c$ as linear combinations of $Q_i$’s and $Q_i$’s, respectively. Specifically,

$$\Phi_A^c = \frac{1}{N} \sum_i \vartheta_{1i} Q_i, \quad \Phi_B^c = \frac{1}{N} \sum_i \vartheta_{2i} Q_i,$$

(7.43)

where we denote

$$\vartheta_{1i} := z \frac{1 - M_{\mu_A}(\Omega_B)}{z m_H(z) + 1} \frac{a_i \text{tr}(\tilde{B}G) - \text{tr}(A\tilde{B}G)}{(a_i - \Omega_B)}, \quad \vartheta_{2i} := z \frac{1 - M_{\mu_B}(\Omega_A)}{z m_H(z) + 1} \frac{b_i \text{tr}(\tilde{A}G) - \text{tr}(B\tilde{A}G)}{b_i - \Omega_A}.$$

(7.44)
Indeed, (7.34) follows from the following decompositions

\[
\Phi^c_A = \frac{m_{\mu_A}(\Omega_B^c)}{(\Omega_B^c m_{\mu_A}(\Omega_B^c) + 1) - z} - \frac{\Omega_A^c}{z} \left[ m_{\mu_A}(\Omega_B^c)(zm_H(z) + 1) - \text{tr}(\bar{B}G)(\Omega_B^c m_{\mu_A}(\Omega_B^c) + 1) \right]
\]

\[
= \frac{1}{(\Omega_B^c m_{\mu_A}(\Omega_B^c) + 1)(zm_H(z) + 1)} \sum_{i,j} \left[ a_j (\bar{B}G)_{jj} - \frac{a_j}{a_i - \Omega_B^c} (\bar{B}G)_{ii} a_j \right]
\]

\[
= \frac{1}{zm_H(z) + 1} \sum_{i,j} a_j \left[ (\bar{B}G)_{jj} \left( \frac{1}{a_i - \Omega_B^c} (\bar{B}G)_{ii} \right) + (\bar{B}G)_{ii} \left( \frac{1}{a_j - \Omega_B^c} \right) \right]
\]

\[
= \frac{z}{N} \frac{1 - M_{\mu_A}(\Omega_B^c)}{zm_H(z) + 1} \sum_{i,j} a_i (\bar{B}G) - \text{tr}(\bar{A}BG) - \text{tr}(\bar{A}BG)
\]

where in the second equality we used (6.1) and the definition in (2.4). Similarly, we have

\[
\Phi^c_B = \frac{m_{\mu_B}(\Omega_A^c)}{(\Omega_A^c m_{\mu_B}(\Omega_A^c) + 1) - z} - \frac{\Omega_A^c}{z} \left[ m_{\mu_B}(\Omega_A^c)(zm_H(z) + 1) - \text{tr}(\bar{G}A)(\Omega_A^c m_{\mu_B}(\Omega_A^c) + 1) \right]
\]

\[
= \frac{1}{(\Omega_A^c m_{\mu_B}(\Omega_A^c) + 1)(zm_H(z) + 1)} \sum_{i,j} \left[ b_j (\bar{G}A)_{jj} - \frac{b_j}{b_i - \Omega_A^c} (\bar{G}A)_{ii} b_j \right]
\]

\[
= \frac{1}{zm_H(z) + 1} \sum_{i,j} b_j \left[ (\bar{G}A)_{jj} \left( \frac{1}{b_i - \Omega_A^c} (\bar{G}A)_{ii} \right) + (\bar{G}A)_{ii} \left( \frac{1}{b_j - \Omega_A^c} \right) \right]
\]

\[
= \frac{z}{N} \frac{1 - M_{\mu_B}(\Omega_A^c)}{zm_H(z) + 1} \sum_{i,j} b_i (\bar{G}A) - \text{tr}(\bar{A}BG) - \text{tr}(\bar{A}BG)
\]

To apply Proposition 7.4, we need to check whether its assumptions are satisfied for our case. Recall (7.33). First, we show that we can choose \( \tilde{\Pi} = \tilde{\Omega}^{1/2} \). By (7.31) and (7.32), we find that

\[
m_H(z) - m_{\mu_A \mathcal{B}_{\mu_B}}(z) = \frac{(zm_H(z) + 1)(zm_{\mu_A \mathcal{B}_{\mu_B}}(z) + 1)}{z} \left[ \frac{1}{zm_{\mu_A \mathcal{B}_{\mu_B}}(z) + 1} - \frac{1}{zm_H(z) + 1} \right]
\]

\[
= \frac{(zm_H(z) + 1)(zm_{\mu_A \mathcal{B}_{\mu_B}}(z) + 1)}{z} \left( \frac{m_{\mu_A \mathcal{B}_{\mu_B}}(z) - zm_H(z) + 1}{zm_H(z) + 1} \right)
\]

\[
= \frac{(zm_H(z) + 1)(zm_{\mu_A \mathcal{B}_{\mu_B}}(z) + 1)}{z} \left( \frac{\Omega_A(z)\Omega_B(z) - \Omega_A^c\Omega_B^c}{z} + \frac{\text{tr}(AG)\text{tr}(\bar{B}G) - \text{tr}(ABG)}{(1 + z \text{tr} G)^2} \right)
\]

Together with Propositions 6.1 6.2 and 7.4, we find that

\[
|m_H(z) - m_{\mu_A \mathcal{B}_{\mu_B}}(z)| < \Lambda(z) + \Psi < \tilde{\Lambda} + \Psi^2.
\]

This yields that

\[
\Pi^2 = \frac{\Im m_H(z)}{N\eta} < \frac{\Im m_{\mu_A \mathcal{B}_{\mu_B}}(z) + \tilde{\Lambda} + \Psi^2}{N\eta} < \Psi^2 \left( \sqrt{\Im m_{\mu_A \mathcal{B}_{\mu_B}}(z) + \tilde{\Lambda}}(|\mathcal{S}_{AB}(z)| + \tilde{\Lambda}) + \Psi^2 \right)
\]
where we used $\text{Im} m_{\mu A \Theta_{\mu B}}(z) \preceq |S_{AB}(z)|$ in the last inequality, which follows from (ii) and (iii) of Proposition 3.1. In other words, we see that
\[
\Pi \prec \Omega^{1/2}. \tag{7.46}
\]
Moreover, by (ii) and (iii) of Proposition 3.1, we conclude that both $\text{Im} m_{\mu A \Theta_{\mu B}}(z)$ and $|S_{AB}(z)|$ are bounded. Consequently, we have that
\[
\Omega^{1/2} \prec \Psi. \tag{7.47}
\]
Using (ii) and (iii) of Proposition 3.1 again, we can conclude that
\[
\frac{1}{N} \sqrt{\eta} \prec \Psi^2 \sqrt{\text{Im} m_{\mu A \Theta_{\mu B}}(z)|S_{AB}(z)|} \prec \Psi^2 \left( \sqrt{\text{Im} m_{\mu A \Theta_{\mu B}}(z) + \Lambda}(|S_{AB}(z) + \Lambda| + \Psi^2) \right).
\]
This implies that
\[
N^{-1/2} \eta^{-1/4} \prec \Omega^{1/2}. \tag{7.48}
\]
By (7.46), (7.47) and (7.48), we have seen that we can choose $\Pi = \Omega^{1/2}$.

Second, we show that the coefficients $\{a_1\}$ and $\{a_2\}$ satisfy Assumption 7.1. Indeed, for $j = 1, 2, \cdots, N$, we can write $\varphi_{j1}$ as
\[
\varphi_{j1} = \frac{z}{\Omega_B^c m_{\mu A} (\Omega_B^c)/(\text{tr}(\bar{B}G))} a_i \text{tr}(\bar{B}G) - \text{tr}(\bar{A}B_G) a_i - \Omega_B^c,
\]
where we used (6.1). Using the definitions of $\Omega_B^c$ in Definition 6.1, we find that $\varphi_{j1}$ can be regarded as a smooth function of $\text{tr}(\bar{B}G)$ and $\text{tr} G$. Then, using chain rule, the general Leibniz rule and Lemma B.3, it is easy to see that $\varphi_{j1}$ satisfies Assumption 7.1. Similar results hold for $\varphi_{j2}$.

Based on the above discussion, we find that both $\Phi_{\mu A}$ and $\Phi_{\mu B}$ satisfy the conditions of Proposition 7.2 with $\Pi = \Omega^{1/2}$. Then Proposition 7.2 implies (7.42). Now we return to (7.40). By (7.41) and (7.42), we find that
\[
\frac{S_{AB}(z)}{z} \Lambda_A + T_A \Lambda_B^2 + O(|\Lambda_A|^3) \prec \hat{\Omega} + \Psi^2 \hat{\Omega} + \Lambda \Psi \Omega^{1/2}. \tag{7.49}
\]
Recall the definition of $\hat{\Omega}$ in (7.34). It is easy to see that
\[
\Lambda \Psi \prec \sqrt{\Lambda \Psi} \prec \Omega^{1/2}.
\]
Together with (7.49), we complete the proof of Proposition 7.4.

We next prove Lemma 7.5. Before stepping into the proof, we have obtained a bound for $Z_1$ and $Z_2$ from the previous discussion. Specially, by using the definitions of $Z_1$ and $Z_2$ in (7.34), a discussion similar to (8.13) and (7.42), we conclude that
\[
Z_1 \prec \Psi \Omega^{1/2}, \ Z_2 \prec \Psi \Omega^{1/2}.
\]

**Proof of Lemma 7.5.** We only focus our proof on $Z_1$ and $Z_2$ can be handled similarly. Recall from (8.35) that $Z_1$ are linear combinations of $\Phi_{\mu A}$ and $\Phi_{\mu B}$. We can write
\[
\mathbb{E} \left[ Z_1^{(p,p)} \right] = \frac{1}{N} \sum_{i=1}^{N} E \left[ \varphi_{i1} Q_1 Z_1^{(p-1,p)} \right] + \frac{z L'_{\mu A}(\Omega_B)}{N} \sum_{i=1}^{N} E \left[ \varphi_{i2} Q_1 Z_1^{(p-1,p)} \right].
\]
Due to similarity, we only state the estimate of the first term on the RHS of the above equation. The proof is again an application of the formula (8.39). By a discussion similar to (7.3) using $\Pi = \Omega^{1/2}$, we find that
\[
\frac{1}{N} \sum_{i=1}^{N} E \left[ \varphi_{i1} Q_1 Z_1^{(p-1,p)} \right] = E \left[ O_\prec(\hat{\Omega}) Z_1^{(p-1,p)} \right] + E \left[ O_\prec(\Psi^2 \bar{\Omega}) Z_1^{(p-2,p)} \right] + E \left[ O_\prec(\Psi^2 \bar{\Omega}) Z_1^{(p-1,p-1)} \right]. \tag{7.50}
\]
Since the estimate of the coefficient in front of the term $3_1^{(p-1,p)}$ matches with that in (7.36), it suffices to improve the estimates of the second and third terms on the RHS of (7.50) in view of (7.47).

In the rest of the proof, we briefly discuss the second term on the RHS of (7.50) and the third term can be estimated similarly. Regarding the coefficients of $3_1^{(p-2,p)}$, by a discussion similar to (7.26), we find that all of them have one of the following forms

$$
\frac{1}{N^2} \sum_i \sum_k (c_i e_k X_i e_i \partial_3 \frac{1}{N^2} \sum_i \sum_k (c_i e_k X_i e_i
$$

where $c_i$ stand for some $O(1)$ factors. By chain rule, we have

$$
\frac{\partial_3 \partial \Phi_A}{\partial g_{ik}} (\Phi_A(\Omega_A^c, \Omega_B^c) + z L_{\mu_A}(\Omega_B) \Phi_B(\Omega_A^c, \Omega_B^c))
$$

= \left( z L_{\mu_A}(\Omega_B) L_{\mu_B}(\Omega_A^c) - \frac{1}{z} \right) \frac{\partial \Omega_A^c}{\partial g_{ik}} + (L_{\mu_A}^c(\Omega_B) - L_{\mu_A}(\Omega_B)) \frac{\partial \Omega_B^c}{\partial g_{ik}}.
$$

From Proposition 3.1 and Assumption 6.2 we find that the coefficients in front of derivatives of $\Omega_A^c$ and $\Omega_B^c$ in the above equation admit the estimates

$$
|z L_{\mu_A}(\Omega_B) L_{\mu_B}(\Omega_A^c) - \frac{1}{z}| \prec |S_{AB}| + \Lambda, \quad |L_{\mu_A}(\Omega_B) - L_{\mu_A}(\Omega_B)| \prec \Lambda.
$$

Since they do not depend on indices $i$ and $k$, we can simply pull them out as scaling factors. As a consequence, in light of (7.43), the remaining weighted sum has the same form as in the second or third estimates in Lemma 3.3 which is $O_{\prec}(\Pi^2 \Psi^2)$. Thus we conclude that both quantities in (7.51) satisfy

$$
\frac{1}{N^2} \sum_i \sum_k (c_i e_k X_i e_i \partial_3 \frac{1}{N^2} \sum_i \sum_k (c_i e_k X_i e_i
$$

where we used the definition of $\hat{\Psi}$ in (7.34) and Proposition 3.1 to obtain

$$
(|S_{AB}| + \Lambda) \Psi^2 \prec \Psi^2 \sqrt{|S_{AB}| + \Lambda} \prec \hat{\Psi} \quad \text{and} \quad \Lambda \Psi^2 \prec \hat{\Psi}.
$$

The proves that the estimate of the coefficient of $3_1^{(p-2,p)}$ is $O_{\prec}(\hat{\Psi}^2)$. This finishes our proof. $\square$

8 Proof of Theorems 2.12 and 2.16 and Proposition 4.1

This section is mainly devoted to the proof of Theorems 2.12 and 2.16 and their linearization Proposition 4.1. The discussion will make use of the weak local law which will be established in Section 8.1

8.1 Weak local laws

In Section 6 we proved estimates for the entries of the resolvents and optimal fluctuation averaging for the linear combinations of them. We also proved the closeness of the subordination functions and their approximates. However, all these results are regarding pointwise control for fixed $z \in D_r(\eta_L, \eta_U)$ and under Assumption 6.2. In this section, we will establish a weak local law without imposing Assumption 6.2 and uniformly in $z \in D_r(\eta_L, \eta_U)$, using a continuous bootstrapping argument. The weak local law will guarantee that Assumption 6.2 holds uniformly for $z \in D_r(\eta_L, \eta_U)$. As a consequence, the results in Section 6 also hold uniformly for $z \in D_r(\eta_L, \eta_U)$.

More specifically, the main result of this subsection is stated in the following proposition.
Proposition 8.1. Suppose that Assumptions \(\text{I.2}\) and \(\text{I.4}\) hold. Let \(\tau > 0\) be a sufficiently small constant and \(\gamma > 0\) be any fixed small constant. Then we have
\[
\Lambda_d(z) < \frac{1}{(N\eta)^{1/3}}, \quad \Lambda(z) \leq \frac{1}{(N\eta)^{1/3}}, \quad \Lambda_T(z) < \Psi(z), \quad \Lambda_o < \frac{1}{(N\eta)^{1/3}}, \quad \Lambda_{\tilde{T}_o} < \Psi(z).
\] (8.1)
uniformly in \(z \in \mathcal{D}_\tau(\eta_L, \eta_U)\). The same statements hold for \(\tilde{\Lambda}_d\) and \(\tilde{\Lambda}_T\).

The proof of Proposition 8.1 will be divided into three steps. In the first step, we prove \(\text{8.1}\) on the global scale such that \(\eta \geq \eta_U\), where \(\eta_U\) is a sufficiently large constant. The key idea for proving this step is to regard \(Q_i\) defined in \([6, \text{Theorem 8.1}]\) as a function of the random unitary matrix \(U\). This strategy has been employed in the study of addition of random matrices in \([7, \text{Theorem 8.1}]\) and local single ring theorem in \([6]\). An advantage of doing so is that we can employ the device of Gromov-Milman concentration inequality. We collect the related results in the following lemma.

Denote \(U(N)\) as the set of \(N \times N\) unitary matrices over \(\mathbb{C}\) and \(SU(N) \subseteq U(N)\) as the set of special unitary matrices defined by
\[
SU(N) := \{U \in U(N) : \det(U) = 1\}.
\]
Recall that both \(U(N)\) and \(SU(N)\) are compact Lie groups with respect to the matrix multiplication.

Lemma 8.2. Let \(f\) be a real-valued Lipschitz continuous function defined on \(U(N)\). We denote its Lipschitz constant \(L_f\) as
\[
L_f := \sup \left\{ \frac{|f(U_1) - f(U_2)|}{\|U_1 - U_2\|_2} : U_1, U_2 \in U(N) \right\}, \quad \|U_1 - U_2\|_2 := \sqrt{\text{Tr}((U_1 - U_2)^*(U_1 - U_2))}.
\]
Moreover, we let \(\nu\) be the Haar measure defined on \(U(N)\) and \(\nu_h\) that on \(SU(N)\). Then for some universal constants \(c, C > 0\), we have that for any constant \(\delta > 0\),
\[
\int_{U(N)} I \left( \left| f(U) - \int_{SU(N)} f(VU)d\nu_h(V) \right| > \delta \right) d\nu(U) \leq C \exp \left( -c \frac{N \delta^2}{L_f^2} \right). \quad (8.2)
\]
Finally, let \(H_N\) be the group of the form \(\{\text{diag}(e^{i\theta}, 1, \cdots, 1) : \theta \in [0, 2\pi]\}\) and \(\nu_h\) be its associated Haar measure in the sense that \(\tilde{\nu}_h\) is uniformly distributed on \([0, 2\pi]\), then we have
\[
\int_{SU(N)} f(VU_1W)d\nu_h(W)d\nu_h(V) = \int_{U(N)} f(U)d\nu(U), \quad \text{for all } U_1 \in U(N). \quad (8.3)
\]

Proof. See Corollary 4.4.28 and Lemma 4.4.29 of \([1]\). \(\square\)

In the second step, we will employ a continuity argument based on the estimates obtained from step one to prove the results hold for each fixed \(z \in \mathcal{D}_\tau(\eta_L, \eta_U)\). To this end, for \(z \in \mathcal{D}_\tau(\eta_L, \eta_U)\) and \(\delta, \delta' \in [0, 1]\), we define events
\[
\Theta(z, \delta, \delta') := \left\{ \Lambda_d(z) \leq \delta, \Lambda(z) \leq \delta, \Lambda_T(z) \leq \delta', \Lambda_{\tilde{T}_o}(z) \leq \delta' \right\}, \quad (8.4)
\]
\[
\Theta_2(z, \delta, \delta', \epsilon) := \Theta(z, \delta, \delta') \cap \left\{ \Lambda(z) \leq N^{-\epsilon} |S_{AB}(z)| \right\}. \quad (8.5)
\]
Moreover, we decompose the domain \(\mathcal{D}_\tau(\eta_L, \eta_U)\) into two disjoint sets,
\[
\mathcal{D}_\tau = \mathcal{D}_\tau(\tau, \eta_L, \eta_U, \epsilon) := \left\{ z \in \mathcal{D}_\tau(\eta_L, \eta_U) : \sqrt{\kappa + \eta} \geq \frac{N^{2\epsilon}}{(N\eta)^{1/3}} \right\},
\]
\[
\mathcal{D}_\tau = \mathcal{D}_\tau(\tau, \eta_L, \eta_U, \epsilon) := \mathcal{D}_\tau(\eta_L, \eta_U) \setminus \mathcal{D}_\tau.
\]
The main technical input for the second step is Lemma 8.3 below, whose proof will be postponed to Appendix \([\text{C}]\). As it will be seen from Lemma 8.3 when we restrict ourselves on some high probability event, it is always possible to gradually improve our estimates.
Lemma 8.3. Suppose that Assumptions 2.2 and 2.4 hold. For any fixed \( z \in D_\tau(\eta_L, \eta_U) \), any \( \epsilon \in (0, \gamma/12) \) and \( D > 0 \), there exists \( N_1(D, \epsilon) \in \mathbb{N} \) and an event \( \Xi(z, D, \epsilon) \) with

\[
\Pr(\Xi(z, D, \epsilon)) \geq 1 - N^{-D}, \quad \text{for all } N \geq N_1(D, \epsilon),
\]

such that the followings hold:

(i) For all \( z \in D_\tau \),

\[
\Theta> \left( z, \frac{N^{3\epsilon}}{(N\eta)^{1/3}}, \frac{N^{3\epsilon}}{\sqrt{N\eta}}, \frac{\epsilon}{10} \right) \cap \Xi(z, D, \epsilon) \subset \Theta> \left( z, \frac{N^{5\epsilon/2}}{(N\eta)^{1/3}}, \frac{N^{5\epsilon/2}}{\sqrt{N\eta}}, \frac{\epsilon}{2} \right).
\]

(ii) For all \( z \in D_\tau \),

\[
\Theta \left( z, \frac{N^{3\epsilon}}{(N\eta)^{1/3}}, \frac{N^{3\epsilon}}{\sqrt{N\eta}} \right) \cap \Xi(z, D, \epsilon) \subset \Theta \left( z, \frac{N^{5\epsilon/2}}{(N\eta)^{1/3}}, \frac{N^{5\epsilon/2}}{\sqrt{N\eta}} \right).
\]

In the third step, we prove the uniformity. With the help of Lemma 8.3, we first extend the results to the whole domain from a discrete lattice of mesh size \( N^{-5} \). Then by using the Lipschitz continuity of subordination functions and the resolvents, we can extend the bounds to the entire domain \( D_\tau(\eta_L, \eta_U) \).

For the rest of this subsection, we prove the weak local law.

Proof of Proposition 8.1. We will follow the three-step strategy to complete our proof. Due to similarity, we only prove \( \Lambda_d, \Lambda \) and \( \Lambda_T \).

Step 1. The goal of this step is to prove the results for \( \eta \geq \eta_U \). A main technical input is the following estimate

\[
|zQ_i| < \frac{1}{\sqrt{N\eta U}},
\]

for all fixed \( z \) with \( \text{Im } z \geq \eta_U \). We mention that (8.7) is slightly weaker than its counterpart, equation (8.35) of [7]. Since \( \eta_U \) is large, (8.7) is already sufficient for our discussion.

To show (8.7), we employ (8.2) for a properly chosen function. Specifically, for \( U \in U(N) \), we let

\[
f(U) = zQ_i(U),
\]

where \( Q_i \) is defined in (6.12) and we regard \( Q_i \) as a function of \( U \). To show \( f(\cdot) \) is Lipschitz for large \( \eta U \), we directly calculate its Lipschitz constant in the spirit of [8] Section 8.1 with a slightly different justification. Indeed, since \( f \) is differentiable, it suffices to bound its derivative in order to obtain the Lipschitz constant.

Consider the adjoint representation of the Lie algebra associated with \( U(N) \) [20]. Note that the natural product of this Lie algebra is the Lie bracket, i.e., the commutator. We now introduce some "curve" based on a bounded \( N \times N \) Hermitian matrix \( X \). Since \( B \) is diagonal, by definition, we see for \( 0 \leq t \leq 1 \)

\[
\text{Ad}_{e^{itX}}UBU^* = e^{itX}UBU^*e^{-itX} = e^{it\text{ad}_X(UBU^*)},
\]

where \( \text{Ad} \) is the adjoint action and \( \text{ad} \) is the derivative of the adjoint action at \( t = 0 \) which turns out to be the Lie bracket. Indeed, by an elementary calculation using Lie bracket, we see that

\[
\frac{d}{dt}(\text{Ad}_{e^{itX}}UBU^*)_{t=0} = \frac{d}{dt}e^{it\text{ad}_X(UBU^*)}(ad_Xe^{it\text{ad}_X(UBU^*)})_{t=0} = ad_XUBU^*.
\]

Furthermore, it is easy to see that

\[
\frac{d}{dt} \left( X_A(e^{itX})B(e^{itX})^*G(e^{itX}U) \right)_{t=0} = X_A(I + GA)\text{ad}_X(UBU^*)G,
\]

(8.9)
where $X_A$ can be $A$, $I$, or $A^2$. Actually, invoking the definition of $Q_i$ in (6.12), using (6.1), the facts $zG + 1 = \tilde{A}G$ and $\text{tr} A = 1$, we can write

$$zQ_i = (A\tilde{B}G)_{ii} \text{tr}(A\tilde{B}G) - \text{tr}(A\tilde{B}G) - \text{tr}(A\tilde{B}GA)(\tilde{B}G)_{ii} + (\tilde{B}G)_{ii}. \quad (8.10)$$

In this sense, it suffices to analyze $(X_A\tilde{B}G)_{ii}$ or $tr(X_A\tilde{B}G)$, where $X_A$ can be $I$, $A$, or $A^2$, in order to study $f(U)$ in (8.4). More specifically, due to the nature of Lie algebra and (8.10), the control of the derivative of $f$ reduces to bound the terms on the RHS of (8.9).

First, for the term of the form $(X_A\tilde{B}G)_{ii}$, by (8.10) and (v) of Assumption 2.4, we see that

$$|e^*_iX_A(I + GA)\text{ad}_X(\tilde{B})Ge_i| \leq \|X(I + GA)^*X_Ae_i\| \|\tilde{B}Ge_i\| + \|\tilde{B}(I + GA)^*X_Ae_i\| \|XGe_i\| \leq C\eta^{-1}_U \|X\| \leq C\eta^{-1}_U \|X\|^2, \quad (8.11)$$

where $C \equiv C(A, B) > 0$ is some constant and $\|X\|^2$ is the Hilbert-Schmidt norm of $X$. Here we used the inequalities $\|G\| \leq \eta^{-1}$ and $\|X\| \leq \|X\|^2$. Second, for the term of the form $\text{tr}(X_A\tilde{B}G)$, by Hölder’s inequality, we have that

$$|\text{tr} X_A(I + GA)\text{ad}_X(\tilde{B})G| \leq \|X_A(I + GA)\| (\text{tr} |X\tilde{B}| + \text{tr} |\tilde{B}X|) \|G\| \leq \frac{C}{N\eta_U} \|X\|^2 \leq \frac{C}{\sqrt{N}\eta_U} \|X\|^2, \quad (8.12)$$

where we used the Cauchy-Schwarz inequality for (8.12). Combining (8.11) and (8.12), we conclude that for some constant $C > 0$,

$$\mathcal{E}_f \leq \frac{C}{\eta_U}. \quad (8.13)$$

Then we show that

$$\int_{SU(N)} f(UV) d\nu_s(V) = 0. \quad (8.14)$$

By (8.3), $B$ is diagonal and the fact $W(\theta)BW(\theta)^* = B$ since $W(\theta) = \text{diag}(e^{i\theta}, 1, \cdots, 1)$, we readily find that

$$\int_{SU(N)} f(UV) d\nu_s(V) = \int f(U) d\nu(U) = z E [Q_i]. \quad (8.15)$$

Furthermore, by Proposition 3.2 and equation (3.25) of [37], we find that

$$E (GA)_{jj} (\tilde{B}G)_{ii} = E (G_{ii}G_{jj}A\tilde{B}). \quad (8.16)$$

By taking the average over the summation of $j$, we obtain $E [Q_i] = 0$. Together with (8.15), we conclude the proof of (8.14).

By (8.13) and (8.14), using (8.2), we have proved the claim (8.7). Armed with this control, we proceed to offer some other useful estimates. We begin with showing that

$$\sup_{z : \text{Im } z \geq \eta_U} |zQ_i| \lesssim N^{-1/2}. \quad (8.16)$$

Indeed, when $|z| \geq \sqrt{N}$, using the definition of the resolvent and (2.11), we find that for some constant $C > 0$,

$$\sup_{|z| \geq \sqrt{N}} \|G\| \leq \sup_{|z| \geq \sqrt{N}} \sup_{i} \{ |z - \lambda_i|^{-1} \} \quad (8.17)$$
\[\leq \sup \{|z-x|^{-1} : a_N b_N + o(1) < x < a_1 b_1 + o(1), |z| \geq \sqrt{N}\}\]
\[\leq \frac{C}{\sqrt{N}}.\]

On the other hand, it is easy to see that there exists a constant \(C > 0\) such that the following statement holds for all fixed \(z_0\) with \(\text{Im}\ z_0 \geq \eta_U\):
\[
\left\{|z_0 Q_i(z_0)| \leq N^{-1/2+\epsilon}\right\} \subset \left\{\sup \left\{|z Q_i(z)| : \text{Im}\ z \geq \eta_U, |z - z_0| \leq N^{-1/2}\right\} \leq C N^{-1/2+\epsilon}\right\}. \tag{8.18}
\]

Since \(G\) is \(\eta_U^{-2}\)-Lipschitz, by \(\textbf{[8.10]}\), \(Q_i\) is also \(\eta_U^{-2}\)-Lipschitz. We take a discrete lattice \(\mathcal{P}\) in \(\{z : \text{Im}\ z \geq \eta_U, |z| \leq \sqrt{N}\}\) with side length \(N^{-1/2}\) and construct a net \(\tilde{\mathcal{P}}\) which is the intersection of \(\mathcal{P}\) and the term on the RHS of \(\textbf{[8.18]}\). Since \(\sup_{z \in \tilde{\mathcal{P}}} |z Q_i(z)| \prec N^{-1/2}\), by Lipschitz continuity, we conclude the proof of \(\textbf{[8.16]}\).

Then we prove the estimate
\[
\sup_i \sup_{|z| > \eta_U} |\Lambda_{di}| \prec N^{-1/2}. \tag{8.19}
\]

To show \(\textbf{[8.19]}\), by \(\textbf{[8.11]}\), we rewrite
\[
\Lambda_{di} = \left\lfloor \frac{a_i}{(1 + z \text{ tr} \ G)(a_i - \Omega_B^c)} \right\rfloor |z Q_i|. \tag{8.20}
\]

In view of \(\textbf{[8.16]}\), it suffices to show that
\[
(1 + z \text{ tr} \ G)(a_i - \Omega_B^c) \prec 1. \tag{8.21}
\]

Since \(\eta_U\) is large enough, for all \(z\) with \(\text{Im}\ z \geq \eta_U\), for some constant \(C > 0\), we have
\[
\left\lfloor 1 + z \text{ tr} \ G \left(1 + \frac{\text{tr}(A \tilde{B})}{z}\right) \right\rfloor = \left|\text{tr}(A \tilde{B}((A \tilde{B} - z)^{-1} + z^{-1}))\right| = \left|\text{tr}((A \tilde{B})^2(z(A \tilde{B} - z))^{-1})\right| \leq \frac{C}{|z|^2}, \tag{8.22}
\]

where in the second equality we used the property of trace and the decomposition \(A \tilde{B} = A \tilde{B} - z + z\), and in the last step we apply Von Neumann’s trace inequality and a discussion similar to \(\textbf{[8.17]}\). Furthermore, by Gromov-Milman concentration inequality \(\textbf{[8.22]}\), we readily obtain
\[
\mathbb{P} \left[\left|\text{tr}(A \tilde{B}) - 1\right| > \delta\right] \leq C \exp(-c(N\delta)^2),
\]
where we used
\[
\mathbb{E} \left[\text{tr}(A \tilde{B}) \right] = \frac{1}{N} \sum_{ij} a_i b_j \mathbb{E} |v_{ij}|^2 = \text{tr} A \text{ tr} B = 1, \quad \left|\text{tr}(A a_{X\tilde{B}})\right| \leq \frac{C}{\sqrt{N}} \|X\|_2.
\]

Since \(v_i\) are uniformly distributed on the unit sphere \(S^N\), we have that \(\mathbb{E} \left[|v_{ij}|^2\right] = 1/N\). Together with \(\textbf{[8.22]}\), we have that
\[
|1 + z \text{ tr} \ G - z^{-1}| \leq C|z|^{-2} + O_{\prec}(\langle |z|N \rangle^{-1}). \tag{8.23}
\]

By a discussion similar to \(\textbf{[8.22]}\), we have
\[
\left|\text{tr}(A G - z^{-1})\right| = \left|\text{tr}(A(G + z^{-1}))\right| = \left|\text{tr}(A^2 \tilde{B}(z(A \tilde{B} - z))^{-1})\right| \leq C|z|^{-2}.
\]

Based on the above calculation, we arrive at
\[
\Omega_B^c = z \frac{\text{tr}(A G)}{1 + z \text{ tr} \ G} = z(1 + O(|z|^{-1}) + O_{\prec}(N^{-1})). \tag{8.24}
\]
Since \( a_i \) is bounded and \( \eta_U \) is large enough, by (8.23) and (8.24), we conclude that
\[
(1 + z \text{ tr } G)(a_i - \Omega_B^*) = 1 + O(|z|^{-1}) + O_{\prec}(N^{-1}).
\]
This proves (8.21) and hence (8.19).

Next, we provide some error bounds regarding the system (7.25). Using (8.23), (8.24), (8.19) and (2.11), we find that
\[
|M_{\mu_B}(z) - M_{\mu_A}(\Omega_B^*)| = \left| \frac{1}{z \text{ tr } G + 1} - \frac{1}{\Omega_B^* m_{\mu_A}(\Omega_B^*) + 1} \right| |z|^2 N^{-1/2},
\]
where in the first step we used the definition (2.4). Recall (6.17) and (6.18). Similarly, we have that
\[
|\Omega_A^* \Omega_B^* - z M_{\mu_B}(z)| = \left| \frac{z}{(1 + z m_{\mu_B}(z))^2} \right| \left| \frac{1}{N} \sum_i z Q_i \right| |z|^3 N^{-1/2}.
\]
Recall the definitions in (7.31). Combining the above two inequalities, by (8.24), we get
\[
|\Phi_A^*| \leq \left| \frac{M_{\mu_A}(\Omega_B^*) - M_{\mu_B}(z)}{\Omega_B^*} \right| + \left| \frac{z M_{\mu_B}(z) - \Omega_A^* \Omega_B^*}{z \Omega_B^*} \right| |z| N^{-1/2}.
\]
On the other hand, by Lemma 3.10 \( |\Phi_A^*| \) also admits the following bound
\[
|\Phi_A^*| \prec |z|^{-1}.
\]
Similarly, we can show that
\[
|\Phi_B^*| \prec |z|^{-1}.
\]
Together with (8.24) and (8.26), by Lemma 3.6 we readily find that uniformly in \( z \in \mathcal{D}_r(\eta_U, \infty) \)
\[
|\Lambda_A| = |\Omega_A^* - \Omega_A(z)| \leq 2(|\Phi_A^*(z)| + |\Phi_B^*(z)|) \prec |z| N^{-1/2}.
\]
(8.26)
Similar results hold for \( \Lambda_B \). Fix \( z \in \mathcal{D}_r(\eta_U, \eta_U) \), say \( z = E + i \eta \), by (8.19), (8.20) and (i) of Proposition 3.1, we conclude that
\[
\sup_i A_{di}(E + i \eta_U) = \sup_i \left| z G_{ii} - \frac{a_i}{a_i - \Omega_B^*} \right| \leq \sup_i \Lambda^*_{ii} + \sup_i \left( \frac{a_i |\Lambda_B^*|}{|a_i - \Omega_B^*| (|a_i - \Omega_B^*| - |\Lambda_B^*|)} \right) \prec N^{-1/2}.
\]
Recall the definition of \( T_i \) in (6.13). Using the trivial bound \( \|G(E + i \eta_U)\| \leq \eta_U^{-1} \), we find that
\[
\Lambda_T(E + i \eta_U) \leq \eta_U^{-1}.
\]
Similarly, we can show that \( \tilde{\Lambda}_T(E + i \eta_U) \leq \eta_U^{-1} \).

Based on the above discussion, since \( \eta_U \) is a sufficiently large constant, we see that Assumption 5.2 holds for \( z = E + i \eta_U \). Then by Proposition 6.3, we have that for \( z = E + i \eta_U \)
\[
\Lambda_T \prec N^{-1/2}, \quad \tilde{\Lambda}_T \prec N^{-1/2}, \quad \Upsilon \prec N^{-1/2}.
\]
Moreover, by (iii) of Proposition 3.1 and (8.26), we have \( \Lambda_A(z) \prec N^{-\epsilon} \sqrt{\eta + \eta_U} \) for \( z = E + i \eta_U \). Quantitively, for any fixed \( E \in \mathbb{R} \),
\[
P \left[ \Theta_\Gamma \left( E + i \eta_U, \frac{N^3 \epsilon}{N \eta_U^{1/2}}, \frac{N^3 \epsilon}{(N \eta_U^{1/2})^{1/2}}, \frac{\epsilon}{10} \right) \right] \geq 1 - N^{-D},
\]
for all \( D > 0 \) and \( N \geq N_2(\epsilon, D) \), where \( N_2(\epsilon, D) \) depends only on \( \epsilon \) and \( D \).

**Step 2.** In this step, with the estimate (8.27) and Lemma 8.3, we control the probability of the “good” events \( \Theta_\Gamma \) for \( z \in \mathcal{D}_\Gamma \), and \( \Theta \) for \( z \in \mathcal{D}_\prec \). Consequently, we can iteratively make \( \text{Im } z \) smaller. More specifically, in this step, we will prove that for the high probability event \( \Xi(\cdot, \cdot, \cdot) \) in (8.6), the following statements hold:
We now prove it. Recall Definition 6.1. Using (6.1) and the fact 
\[ d = \frac{N^{5r/2}}{(N\eta)^{1/3}} \frac{N^{5r/2}}{\sqrt{N\eta}} \epsilon \frac{\epsilon}{2} \] 
where we used (iv) of Proposition 3.1. Second, we claim the following result
\[ \Lambda \leq \delta \] 
for some constant \( \delta \) and \( \delta' \).

We decompose the domain \( D \) into \( D_\geq \) and \( D_\leq \) as in \( \Theta \), we need to keep track of the event \( \Lambda \leq N^{-\epsilon/2} S_{AB} \) in order to apply (i) of Lemma 8.3; see equation (8.33) for more details.

We start with discussing the event \( \Theta \). First, for generic values \( z, \delta \) and \( \delta' \), we claim that on the event \( \Theta(z, \delta, \delta') \) defined in (8.6), there exists some constant \( C > 0 \),
\[ \Lambda(z + w) \leq \delta + CN^{-3}, \quad \Lambda T(z + w) \leq \delta' + CN^{-3}, \] 
for all \( w \in \mathbb{C} \) with \( |w| \leq 2N^{-5} \), and \( z + w \in D_\tau(\eta \nu, \nu) \). Indeed, the above estimates follow from mean value theorem and
\[ \left\| \frac{d}{dz} (zG + 1) \right\| \leq \eta^{-2} \leq N^2, \quad |\Omega_A(z)| \leq \frac{C}{\sqrt{\kappa + \eta}} \leq N^{1/2}, \quad |\Omega_B(z)| \leq \frac{C}{\sqrt{\kappa + \eta}} \leq N^{1/2}, \] 
where we used (iv) of Proposition 3.1. Second, we claim the following result
\[ \Lambda(z + w) \leq \delta' + CN^{-3}. \] 
We now prove it. Recall Definition 6.1. Using (6.1) and the fact \( \frac{dG}{dz} = G^2 \), we find that
\[ \frac{d}{dz} \Omega_A(z) = \frac{\text{tr}(\overline{B}ABG^2) - \text{tr}(\overline{B}ABG) \text{tr}(ABG^2)}{\text{tr}(ABG^2)}. \] 
For some constant \( C > 0 \), by (6.11), (i) of Proposition 5.1 and the definition of \( \Theta(z, \delta, \delta') \), we have
\[ |\text{tr}(\overline{B}ABG)| = |\text{tr}(\overline{B}(zG + 1))| \leq 1 + \left| \frac{Cz}{N} \sum_{i} \frac{1}{\alpha_i - \Omega_B} \right| + C|z|\delta \leq C, \] 
and
\[ |\text{tr}(\overline{B}ABG)| \geq |zm_{\mu\alpha}g_{\mu\beta}(z) + 1| - \delta. \] 
Combining the above discussions, when \( \delta \ll 1 \), we obtain that
\[ \left| \frac{d}{dz} \Omega_A(z) \right| \leq CN^2. \] 
Similar result hold for \( \Omega_B^\epsilon \). Then we can complete the proof of (8.31) with mean value theorem. ArmEd with (8.30) and (8.31), we immediately see that
\[ \Theta \left( \frac{z}{(N\eta)^{1/3}}, \frac{N^{5r/2}}{\sqrt{N\eta}} \right) \subset \bigcup_{|w| \leq N^{-5}} \Theta \left( \frac{z + w}{(N\eta)^{1/3}} + CN^{-3}, \frac{N^{5r/2}}{\sqrt{N\eta}} + CN^{-3} \right) \subset \bigcup_{|w| \leq N^{-5}} \Theta \left( \frac{z + w}{(N\eta)^{1/3}}, \frac{N^{3r}}{\sqrt{N\eta}} \right), \] 
(8.32)
We next briefly discuss the event $\Theta_>$ due to similarity. On the event $\Theta_> \left( z, \frac{N^{5/2}}{(N\eta)^{1/3}}, \frac{N^{5/2}}{\sqrt{N\eta}}, \frac{\epsilon}{10} \right)$, by a discussion similar to (8.30) and (iv) of Proposition 3.1, we have that for some constants $c, C > 0$,
\[
\Lambda(z - N^{-5}i) \leq \Lambda(z) + CN^{-3} \leq N^{-1/2} |\mathcal{S}_{AB}(z)| + CN^{-3} \\
\leq CN^{-1/2}(\kappa + \eta)^{-1/2} + CN^{-3} \leq cN^{-10}(\kappa + \eta - N^{-5})^{-1/2} \leq N^{-10} |\mathcal{S}_{AB}(z - N^{-5}i)|.
\]
Consequently, we have that
\[
\Theta_> \left( z, \frac{N^{5/2}}{(N\eta)^{1/3}}, \frac{N^{5/2}}{\sqrt{N\eta}}, \frac{\epsilon}{2} \right) \subset \bigcap_{|w| \leq N^{-5}} \Theta_> \left( z + w, \frac{N^{3\epsilon}}{(N\eta)^{1/3}}, \frac{N^{3\epsilon}}{\sqrt{N\eta}}, \frac{\epsilon}{10} \right).
\]
By Lemma 8.3 and taking $w = -N^{-5}i$ in (8.32) and (8.34), we have proved (8.28) and (8.29).

**Step 3.** In this step, we use a continuity argument to extend the bound first to a lattice of mesh size $N^{-5}$ and then the whole domain $\mathcal{D}(\eta_L, \eta_U)$. To simplify the notations, we denote
\[
\mathcal{P} := \mathcal{D}(\eta_L, \eta_U) \cap (N^{-5}\mathbb{Z})^2, \quad \mathcal{P}_> := \mathcal{P} \cap (N^{-5}\mathbb{Z})^2,
\]
\[
\mathcal{P}_\leq := \mathcal{P} \cap (N^{-5}\mathbb{Z})^2, \quad \mathcal{P}_E := [E_+ - \tau, \tau^{-1}] \cap N^{-5}\mathbb{Z}.
\]
Repeatedly applying (8.28) and (8.29), we have
\[
\bigcap_{z \in \mathcal{P}} \Xi(z, D, \epsilon) \cap \bigcap_{E \in \mathcal{P}_E} \Theta_> \left( E + i\eta_U, \frac{N^{5\epsilon/2}}{(N\eta_U)^{1/3}}, \frac{N^{5\epsilon/2}}{(N\eta_U)^{1/2}}, \frac{\epsilon}{2} \right) \subset \bigcap_{z \in \mathcal{P}_>} \Theta_> \left( E + i\eta_U, \frac{N^{5\epsilon/2}}{(N\eta_U)^{1/3}}, \frac{N^{5\epsilon/2}}{(N\eta_U)^{1/2}}, \frac{\epsilon}{2} \right) \cap \bigcap_{z \in \mathcal{P}_\leq} \Theta_> \left( E + i\eta_U, \frac{N^{3\epsilon}}{(N\eta_U)^{1/3}}, \frac{N^{3\epsilon}}{(N\eta_U)^{1/2}}, \frac{\epsilon}{10} \right)
\]
where in the third step we used (8.32) and (8.34). Moreover, by Lemma 8.3 and (8.27), when $N \geq \max(N_1(D, \epsilon), N_2(D, \epsilon))$, the probability of the first event of (8.35) is at least
\[
1 - \sum_{E \in \mathcal{P}_E} N^{-D} - \sum_{z \in \mathcal{P}} N^{-D} \geq 1 - CN^{10-D}.
\]
Since $D$ is arbitrary, we can prove for the discrete lattice by choosing $D > 10$ large enough. Finally, by the Lipschitz continuity of the resolvent as demonstrated in the discussion below (8.38) and the subordination functions in (iv) of Proposition 3.1, we can extend all the bounds from the discrete lattice to $\mathcal{D}(\eta_L, \eta_U)$. This concludes the proof of Proposition 8.1.

\[\Box\]

### 8.2 Strong local law: proof of Theorem 2.12

In this section, using the weak local law Proposition 8.1 as an initial input, we prove Theorem 2.12. Indeed, the proof follows from the same three-step strategy as mentioned in the proof of Proposition 8.1 except that we have better estimates. We first prepare some technical ingredients. The first ingredient is Lemma 8.3 below, which provides a device for us to gradually improve the estimate of $\Lambda(z)$ with fixed $\text{Im} z$. Its proof will be put into Appendix C.

**Lemma 8.4.** Let $\tilde{\Lambda}(z)$ be a deterministic control parameter such that $(N\eta)^{-1} \leq \tilde{\Lambda}(z) \leq N^{-\gamma/4}$. Moreover, denote $\Xi(z)$ as an event on which $\Lambda(z) \leq \tilde{\Lambda}(z)$ holds. Then for any fixed $\epsilon_0 \in (0, \gamma/12)$ and large $D > 0$, there exists $N_0 \equiv N_0(\epsilon_0, D)$ such that the followings hold for all $N \geq N_0$ and $z \in \mathcal{D}(\eta_L, \eta_U)$:
(i) If \( \sqrt{k + \eta} > N^{-\epsilon_0} \Lambda \), there exists a sufficiently large constant \( K_0 > 0 \), such that

\[
P \left[ \Xi(z) \cap \left[ \Lambda \left( \frac{|S_{AB}|}{K_0} \right) > N^{-2\epsilon_0} \Lambda + \frac{N^{7\epsilon_0/5}}{N\eta} \right] \right] \leq N^{-D}, \quad \epsilon = A, B.
\]

(ii) If \( \sqrt{k + \eta} \leq N^{-\epsilon_0} \Lambda \), we have

\[
P \left[ \Xi(z) \cap \left[ |A| > N^{-\epsilon_0} \Lambda + \frac{N^{7\epsilon_0/5}}{N\eta} \right] \right] \leq N^{-D}.
\]

Our second input, Lemma 8.5, is analogous to Lemma 8.3. It serves as the key input to gradually extend the bounds to the entire domain \( \mathcal{D}_\tau(\eta_L, \eta_U) \). Its proof will be given in Appendix C. To this end, we define the domains \( \bar{\mathcal{D}}_\tau, \mathcal{D}_\tau \) and events \( \Theta, \Theta_\tau \) as follows:

\[
\bar{\mathcal{D}}_\tau := \left\{ z \in \mathcal{D}_\tau(\eta_L, \eta_U) : \sqrt{k + \eta} > \frac{N^{2\epsilon_0}}{N\eta} \right\}, \quad \mathcal{D}_\tau := \left\{ z \in \mathcal{D}_\tau(\eta_L, \eta_U) : \sqrt{k + \eta} > \frac{N^{2\epsilon_0}}{N\eta} \right\},
\]

\[
\Theta(z, \delta) := \left\{ \Lambda \leq \delta \right\}, \quad \Theta_\tau(z, \delta, \epsilon') := \Theta(z, \delta) \cap \left\{ \Lambda(z) \leq N^{\epsilon'} |S_{AB}| \right\}.
\]

**Lemma 8.5.** For any fixed \( \epsilon_0 \in (0, \gamma/12) \) and large \( D > 0 \), there exists \( N_0 = N_0(\epsilon_0, D) \) such that the followings hold for all \( N \geq N_0 \) and \( z \in \mathcal{D}_\tau(\eta_L, \eta_U) \):

\[
P \left[ \Theta_\tau \left( z, \frac{N^{3\epsilon_0}}{N\eta}, \frac{\epsilon_0}{10} \right) \right] \leq N^{-D}, \quad \text{for } z \in \bar{\mathcal{D}}_\tau,
\]

\[
P \left[ \Theta \left( z, \frac{N^{3\epsilon_0}}{N\eta} \right) \right] \leq N^{-D}, \quad \text{for } z \in \mathcal{D}_\tau.
\]

Then we prove Theorem 2.12 using Lemmas 8.5 and 8.3.

**Proof of Theorem 2.12** First, we prove that

\[
\Lambda < \frac{1}{N\eta}.
\]

holds uniformly in \( z \in \mathcal{D}_\tau(\eta_L, \eta_U) \). As we mentioned earlier, the proof follows the same three-step proof strategy of Proposition 8.1. We focus on explaining step 1, which differs the most from its counterpart of the proof of Proposition 8.1. Analogous to (8.27), step 1 will prove that for fixed \( \text{Im } z = \eta_U \),

\[
\inf_{E \in [E_{\tau - r}, \tau - 1]} P \left[ \Theta_\tau \left( E + i\eta_U, \frac{N^{3\epsilon_0}}{N\eta_U}, \frac{\epsilon}{10} \right) \right] \geq 1 - N^{-D}.
\]

To see (8.39), on one hand, we notice that by Proposition 8.1, we can initially choose \( \Lambda(z) = N^{3\epsilon_0}(\eta_U)^{-1/3} \) in Lemma 8.3. On the other hand, we will repeatedly apply Lemma 8.3 to gradually improve the control parameter \( \Lambda(z) \) until it reaches the bound \( (N\eta_U)^{-1}(\text{up to a factor of } N^{\epsilon_0}) \). To facilitate our discussion, we denote

\[
\Lambda_1(z) := \frac{N^{3\epsilon_0}}{(N\eta_U)^{1/3}}, \quad \Lambda_k(z) := N^{(4-k)\epsilon_0} \left( \frac{1}{(N\eta_U)^{1/3}} + \frac{N^{2\epsilon_0}}{N\eta_U} \right), \quad \text{for } k \in \left\{ 2, \left[ \frac{2}{3\epsilon_0} + 4 \right] \right\}.
\]

Within the above preparation, we will prove (8.39) by induction in the sense that for each \( k \), there exists an \( N_k \equiv N_k(\epsilon_0, D) \) so that

\[
P \left[ \Lambda(E + i\eta_U) \leq \Lambda_k(E + i\eta_U) \right] \geq 1 - kN^{-D}.
\]

(8.40)
for all \( N \geq N_k \) and uniformly in \( E \in [E_+ - \tau, \tau^{-1}] \). Clearly, we have that \((N\eta)^{-1} \leq \hat{\Lambda}_k(z) \leq N^{-\gamma/4}\) for all \( k \). Therefore, Lemma 8.4 is applicable with all the choices \( \tilde{\Lambda} = \hat{\Lambda}_k \). Since \( \sqrt{\kappa + \eta_U} > N^{-\epsilon_0} \Lambda_k \), by Lemma 8.4 we have

\[
P\left[ \left( \Lambda(E + i\eta_U) \leq \hat{\Lambda}_k(E + i\eta_U) \right) \cap \left( \left( \Lambda \leq \frac{|S_{AB}|}{K_0} \right) \right) \right] \leq N^{-D},
\]

(8.41)

for all \( N \geq N_k' \equiv N_k'(\epsilon_0, D) \) and uniformly in \( E \in [E_+ - \tau, \tau^{-1}] \). By (iii) of Proposition 3.1 we find that for some constant \( c > 0 \),

\[
\frac{|S_{AB}|}{K_0} \geq c\sqrt{\kappa + \eta_U} \geq \hat{\Lambda}_k \geq \Lambda,
\]
on the event \([\Lambda \leq \hat{\Lambda}_k] \). Therefore, we can safely remove the factor of indicator function in (8.41). Consequently, (8.40) leads to

\[
P\left[ |\Lambda_i| > N^{-2\epsilon_0} \hat{\Lambda}_k + \frac{N^7\epsilon_0/5}{N\eta_U} \right]
\]

\[
\leq P[\Lambda > \hat{\Lambda}_k] + P\left[ \left( \left( \Lambda \leq \frac{|S_{AB}|}{K_0} \right) \right) \cap \left( \left( |\Lambda_i| > N^{-2\epsilon_0} \hat{\Lambda}_k + \frac{N^7\epsilon_0/5}{N\eta_U} \right) \right) \right]
\]

\[
\leq (k + 1)N^{-D},
\]

whenever \( N \geq \max\{N'_k, N_k\} \) and \( z = E + i\eta_U \), uniformly in \( E \in [E_+ - \tau, \tau^{-1}] \). Together with the fact

\[
N^{-2\epsilon_0} \hat{\Lambda}_k + \frac{N^7\epsilon_0/5}{N\eta_U} = N^{-\epsilon_0} \hat{\Lambda}_{k+1} + \frac{N^7\epsilon_0/5 - N^{-\epsilon_0}}{N\eta_U} \leq \hat{\Lambda}_k,
\]

we have proved that for (8.40), the same inequality holds for the case \( k + 1 \) if the case \( k \) holds true. Recall that by Proposition 8.1 we have that (8.40) holds for \( k = 1 \). By induction, (8.40) holds for all fixed \( k \). In particular, taking \( k = \lceil 2/(3\epsilon_0) \rceil + 4 \), we have

\[
\hat{\Lambda}_k \leq \frac{N^{4\epsilon_0 - k\epsilon_0 + 2/3} + N^{2\epsilon_0}}{N\eta_U} \leq \frac{N^{3\epsilon_0}}{N\eta_U}.
\]

Since by (iii) of Proposition 8.1 \( \Lambda \leq \hat{\Lambda}_k \) implies \( \Lambda \leq N^{-\epsilon_0/10} |S_{AB}| \) for \( \eta = \eta_U \), we have proved (8.39). Steps 2 and 3 of the proof of (8.39) follows analogously as the counterparts of the proof of Proposition 8.1 using Lemma 8.5 and (8.39), we omit the details here.

Second, with (8.33) and the local laws, we can see that the bounds in Theorem 2.12 hold. In detail, with the local law in Proposition 8.1 we see that Assumption 6.2 and the assumption (6.81) hold uniformly in \( z \in \mathcal{D}_r(\eta_U, \eta_U) \). Consequently, Propositions 6.3 and (6.7) hold uniformly in \( z \in \mathcal{D}_r(\eta_U, \eta_U) \). Then the bounds for the off-diagonal entries, i.e., (2.17) and (2.19), follow immediately. For the diagonal entries, we focus on explaining (2.16). Recall (6.11). We see that for any deterministic \( v_1, \cdots, v_N \in \mathbb{C} \), we can write

\[
\frac{1}{N} \sum_{i=1}^N v_i \left( zG_{ii} + 1 - \frac{a_i}{a_i - \Omega_B^c(z)} \right) = \frac{1}{N} \sum_{i=1}^N \frac{zv_ia_i}{(1 + z\text{tr} G)(a_i - \Omega_B^c(z))}Q_i.
\]

Regarding \( \frac{zv_ia_i}{(1 + z\text{tr} G)(a_i - \Omega_B^c(z))} \) as the random coefficients \( d_i \) in (7.2), it is easy to see that the conditions in (7.2) hold. Hence, by Proposition 7.2 and the weak local law, we find that

\[
\left| \frac{1}{N} \sum_{i=1}^N v_i \left( zG_{ii} + 1 - \frac{a_i}{a_i - \Omega_B^c(z)} \right) \right| \prec \Psi \Pi.
\]

Together with (8.38), we conclude the proof. \( \square \)
8.3 Proof of Theorem 2.16 and Proposition 4.1

In this section, we prove the local laws which are outside the bulk spectrum. We first prove Proposition 4.1.

Proof of Proposition 4.1. We first prove the first part of the results when \( z \in D_\tau(\eta_L, \eta_U) \).

Recall (6.76), (6.77), and (6.78). Along the proof of Proposition 6.3, we have proved that under Assumption 6.2

\[
|P_i| \prec \Pi_i, \quad |K_i| \prec \Pi_i.
\]  

(8.42)

Applying the same argument to the off-diagonal elements defined in (6.21), we can obtain that when \( i \neq j \)

\[
|P_{ij}| \prec \Pi_i + \Pi_j, \quad |K_{ij}| \prec \Pi_i + \Pi_j
\]  

(8.43)

Since we already have proved in Proposition 8.1 that Assumption 6.2 and the assumption (6.81) hold uniformly in \( z \in D_\tau(\eta_L, \eta_U) \), (8.42) indeed holds uniformly in \( z \in D_\tau(\eta_L, \eta_U) \). Moreover, applying Proposition 7.2 to the weights \( d_i \equiv 1, i = 1, 2, \ldots, N \), and \( \hat{\Pi} = \Psi \), we see that \( \Upsilon \prec \Psi^2 \). Therefore, by (ii) of Proposition 3.1, we conclude that for \( i \in I_1 \)

\[
|G_{ii} - \Theta_{ii}| \prec |Q_i| \prec |P_i| + \frac{1}{N\eta}
\]  

(8.44)

\[
\prec \sqrt{\frac{\text{Im} G_{ii} + \text{Im} G_{ii}}{N\eta} + \frac{1}{N\eta}} \prec \sqrt{\frac{\text{Im} m_{\mu,\alpha} \otimes \mu_B}{N\eta} + (N\eta)^{-1}} + \frac{1}{N\eta}
\]

\[
\prec \sqrt{\frac{\text{Im} m_{\mu,\alpha} \otimes \mu_B}{N\eta} + \frac{1}{N\eta}}.
\]

Similarly, for \( i \neq j \in I_1 \), we get

\[
|G_{ij}| \prec |Q_{ij}| \prec \sqrt{\frac{\text{Im} m_{\mu,\alpha} \otimes \mu_B}{N\eta} + \frac{1}{N\eta}}.
\]  

(8.45)

Applying the exact same reasoning to \( G \) proves the same bounds hold for \( |(G - \Theta)_{\mu\nu}| \) with \( \mu, \nu \in I_2 \).

It only remains to consider the case when \( i \in I_1 \) and \( \mu \in I_2 \). Here we use the fact that \( Y^* G = \sqrt{B} U G \sqrt{A} \) to get

\[
G_{i\mu} = z^{-1/2}(Y^* \tilde{G})_{i\mu} = z^{-1/2} \sqrt{b_i \alpha_{\mu}} (U^* G)_{i\mu} = z^{-1/2} \sqrt{b_i \alpha_{\mu}} e^{i\theta_i} T_{i\mu},
\]

where we denote \( T_{i\mu} = T_{\mu-i} \). On the other hand, using the definitions in (6.21), we have

\[
K_{ij} = (1 + b_i \text{tr}(GA) - \text{tr}(GAB)) T_{ij} + Q_{ij}.
\]

Recall from (6.33) that

\[
1 + b_i \text{tr}(GA) - \text{tr}(GAB) = (z m_{\mu,\alpha} \otimes \mu_B) (z) + \frac{\Omega_B(z)}{z} (b_i - \Omega_A(z)) + \frac{1}{N\eta} \sim 1.
\]

We readily obtain that

\[
|G_{i\mu}| \leq CT_{i\mu} \prec |K_{i\mu}| + |Q_{i\mu}| \prec \sqrt{\frac{\text{Im} m_{\mu,\alpha} \otimes \mu_B}{N\eta} + \frac{1}{N\eta}},
\]

where \( T_{i\mu} = T_{i(\mu-i)} \) and similarly for \( K_{i\mu} \) and \( Q_{i\mu} \). This concludes the first part of Proposition 4.1.

The second part of the results follow from (ii) of Proposition 3.1 and the first part of the results. The calculation is standard in the literature of Random Matrix Theory, for instance, see Theorem 3.12 and its proof in [12]. We omit the details here. \( \Box \)
Proof of Theorem 2.16. By Proposition 4.1, we see that the control of the off-diagonal entries, i.e., part (2) of the results have been proved in Proposition 4.1. It remains to prove part (1). We fix an arbitrary chosen \( \epsilon \in (0, \tau/100) \) and consider the event \( \Xi \) on which we have

\[
\sup_{z \in D^+(\eta_L, \nu)} \eta |m_H(z) - m_{\mu_A g_{\mu_B}}(z)| \leq N^{-1+\epsilon/2}, \quad \max_{1 \leq i \leq N/3} i^{1/3} |\lambda_i - \gamma_i| \leq N^{-2/3+\epsilon}, \tag{8.46}
\]

where in the proof we choose \( \eta_L = N^{-1+\epsilon} \). By Theorems 2.12 and 2.13, we have \( P(\Xi) \geq 1 - N^{-D} \) for any large \( D > 0 \). For all \( z_0 = E_0 + i\eta_0 \in D(\eta_L) \) with \( 4\eta_0 \leq \kappa_0 = |E_0 - E_+| \), we consider a counter-clockwise square contour \( C(z_0) \) with side length \( \kappa_0 \) and (bary)center \( z_0 \). Then, on the event \( \Xi \), by Cauchy’s theorem, we have

\[
m_H(z_0) - m_{\mu_A g_{\mu_B}}(z_0) = \left( \int_{C_>(z_0)} + \int_{C_<(z_0)} \right) \frac{m_H(z) - m_{\mu_A g_{\mu_B}}(z)}{z - z_0} \mathrm{d}z,
\]

where \( C_>(z_0) = C(z_0) \cap \{ z : |z| > \eta_L \} \) and \( C_<(z_0) = C(z_0) \cap \{ z : |z| \leq \eta_L \} \). On the contour \( C_>(z_0) \), we use the first bound in (8.46) to get that for some constant \( C > 0 \)

\[
\left| \int_{C_>(z_0)} \frac{m_H(z) - m_{\mu_A g_{\mu_B}}(z)}{z - z_0} \mathrm{d}z \right| \leq N^{-1+\epsilon/2} \frac{2}{\kappa_0} \left| \int_{C_>(z_0)} \frac{1}{\Im z} \mathrm{d}z \right| \leq N^{-1+\epsilon/2} \frac{4}{\kappa_0} \left( \frac{\kappa_0}{\eta_0 + \kappa_0/2} + \log \left( \frac{\eta_0 + \kappa_0/2}{\eta_L} \right) \right) \leq C \frac{N^{-1+\epsilon/2} \log N}{\kappa_0 + \eta_0} \leq N^{-1+\epsilon}. \tag{8.47}
\]

On the other hand, for \( z \) on the other contour \( C_<(z_0) \), we use

\[
|m_H(z)| \leq \frac{1}{N} \sum_i \frac{1}{\lambda_i - \eta_0 - \gamma_i - \epsilon^{-1/3}} \leq \frac{1}{N} \sum_{i \leq N/3} \frac{1}{\lambda_i - \eta_0} + \frac{1}{N} \sum_{i > N/3} \frac{1}{\lambda_i - \eta_0 - \gamma_i - \epsilon^{N/3}} \leq C,
\]

where in the second step we used the second bound in (8.46) and in the third step we used the fact that \( \gamma_i \sim i^{2/3} N^{-2/3} \). Following the same argument and using Lemma 5.2, we get \( |m_{\mu_A g_{\mu_B}}(z)| \leq C \). Using the above bounds, we get

\[
\left| \int_{C_<(z_0)} \frac{m_H(z) - m_{\mu_A g_{\mu_B}}(z)}{z - z_0} \mathrm{d}z \right| \leq C \frac{\eta_L}{\kappa_0} = C \frac{N^\epsilon}{N \kappa_0} \leq C \frac{N^\epsilon}{\kappa_0 + \eta_0} N^{-1+\epsilon}. \tag{8.48}
\]

Combining (8.47) and (8.48), we conclude our proof.

\[\square\]

A Collection of derivatives

In this section, we collect some results involving derivatives. They can be easily checked by elementary calculation.

Lemma A.1. Recall the notations in (2.3), (6.7), (6.8), (6.15) and (6.16). We have the following identities

\[
\frac{\partial G}{\partial g_{ik}} = -GA \frac{\partial B}{\partial g_{ik}} G = -GA \left( \frac{\partial R_i U^{(i)} B(U^{(i)})^* R_i}{\partial g_{ik}} \right) G = -GA \left( \frac{\partial R_i}{\partial g_{ik}} \frac{B^{(i)} R_i}{\partial g_{ik}} + \frac{R_i}{\partial g_{ik}} \frac{\partial B^{(i)} R_i}{\partial g_{ik}} \right) G, \tag{A.1}
\]

\[
\frac{\partial R_i}{\partial g_{ik}} = -\frac{\partial (\ell^2)}{\partial g_{ik}} (e_i + h_i)(e_i + h_i)^* - \ell^2 \left( \frac{\partial h_i}{\partial g_{ik}} e_i + e_i \frac{\partial h_i}{\partial g_{ik}} + \frac{\partial h_i}{\partial g_{ik}} h_i + h_i \frac{\partial h_i}{\partial g_{ik}} \right), \tag{A.2}
\]

\[
\frac{\partial h_i}{\partial g_{ik}} = \frac{1}{\|g_i\|} \frac{\partial g_i}{\partial g_{ik}} + \frac{\partial \|g_i\|^{-1}}{\partial g_{ik}} g_i \|g_i\|^{-1} e_k - \|g_i\|^{-1} e_k \frac{\partial g_i}{\partial g_{ik}} \|g_i\|^{-1} e_k = \|g_i\|^{-1} (e_k - \overline{h}_{ik} h_k), \tag{A.3}
\]

68
\[
\frac{\partial h_i^*}{\partial g_{ik}} = -\frac{1}{2} \|g_i\|^{-3} g_{ik} g_i^* = -\frac{1}{2} \|g_i\|^{-1} \mathbf{t}_{ik} h_i^*,
\]
(A.4)
\[
\frac{\partial \ell_i^2}{\partial g_{ik}} = 2 \frac{\partial \|e_i + h_i\|^2}{\partial g_{ik}} = -2 \|e_i + h_i\|^{-2} \frac{\partial \|e_i + h_i\|^2}{\partial g_{ik}} = -2 \|e_i + h_i\|^{-4} \left( \frac{\partial h_i^* e_i}{\partial g_{ik}} + \frac{\partial e_i^* h_i}{\partial g_{ik}} \right),
\]
(B.5)
\[
\frac{\partial P_i}{\partial g_{ik}} = e_i^* \frac{\partial G}{\partial g_{ik}} e_i \text{tr}(A \tilde{B} G) + z G_{ii} \frac{\partial G}{\partial g_{ik}} \text{tr}(A) (\tilde{B} G)_{ii} - \text{tr}(G A) e_i^* A^{-1} \frac{\partial G}{\partial g_{ik}} e_i \\
+ (e_i^* \frac{\partial G}{\partial g_{ik}} e_i + \frac{\partial T_i}{\partial g_{ik}}) \text{tr}(G_{ii} + T_i) \frac{\partial \mathcal{Y}}{\partial g_{ik}},
\]
(A.6)
\[
\frac{\partial \mathcal{Y}}{\partial g_{ik}} = z \left( \text{tr}(2 z G + 1) \frac{\partial G}{\partial g_{ik}} - \text{tr} \left( \frac{\partial G}{\partial g_{ik}} A (\tilde{B} G)_{ii} \right) - z \text{tr}(G A) \frac{\partial A^{-1}}{\partial g_{ik}} \right),
\]
\[
\frac{\partial K_i}{\partial g_{ik}} = \frac{\partial T_i}{\partial g_{ik}} + \text{tr} \left( \frac{\partial G}{\partial g_{ik}} A (b_i T_i + (\tilde{B} G)_{ii}) \right) + \text{tr}(G A) \left( \frac{\partial b_i T_i}{\partial g_{ik}} + \frac{z}{a_i} e_i \frac{\partial G}{\partial g_{ik}} e_i \right) \\
- z \text{tr} \left( \frac{\partial G}{\partial g_{ik}} (G_{ii} + T_i) - \text{tr}(G \tilde{B}) \right) \left( e_i \frac{\partial G}{\partial g_{ik}} e_i + \frac{\partial T_i}{\partial g_{ik}} \right).
\]
(A.7)

**B Some auxiliary lemmas**

In this section, we prove some auxiliary lemmas.

**B.1 Large deviation inequalities**

In this subsection, we provide some large deviation type controls.

**Lemma B.1** (Large deviation estimates). Let \( X \) be an \( N \times N \) complex-valued deterministic matrix and let \( y \in \mathbb{C}^N \) be a deterministic complex vector. For a real or complex Gaussian random vector \( g \in \mathbb{C}^N \) with covariance matrix \( \sigma^2 1_N \), we have
\[
\|y^* g\| \prec \sigma \|y\|, \quad \|g^* X g - \sigma^2 N \text{ tr } X\| \prec \sigma^2 \|X\|_2.
\]
(B.1)

**Proof.** See Lemma A.1 of [1]. \( \square \)

**Lemma B.2.** Let \( X_i \) be \( \tilde{B}^{(i)} \) or \( I, \) \( X_A \) be \( A \) or \( A^{-1}, \) and \( D = (d_i) \) be a random diagonal matrix with \( \|D\| \prec 1. \) Recall \([6,52]\) and \([6,10]\). Under the assumption of \([6,23]\), for each fixed \( z \in \mathcal{D}, (\eta_L, \eta_U), \) we have
\[
\frac{1}{N} \sum_{k}^{(i)} e_k^* X_i \Delta_G(i,k) e_i \prec \Pi_i^2, \quad \frac{1}{N} \sum_{k}^{(i)} e_k^* X_i G e_i \text{tr}(D X_A \Delta_G(i,k)) \prec \Pi_i^2 \Psi^2,
\]
(B.2)
\[
\frac{1}{N} \sum_{k}^{(i)} e_k^* X_i g_i \text{tr}(D X_A \Delta_G(i,k)) \prec \Pi_i^2 \Psi^2, \quad \frac{1}{N} \sum_{k}^{(i)} e_k^* X_i G e_i e_i^* X_A \Delta_G(i,k) e_i \prec \Pi_i^2,
\]
(B.3)
\[
\frac{1}{N} \sum_{k}^{(i)} e_k^* X_i G e_i h_i^* \Delta_G(i,k) e_i \prec \Pi_i^2.
\]
(B.4)
Proof. Recall the definitions in (B.32) and (B.34). It is easy to see that every term in $\Delta_G(i,k)$ belongs to one of the following forms

$$d_i \tilde{h}_{ik} G \alpha_i \beta_i^* \tilde{B}^{(i)} R_i G \text{ or } d_i \tilde{h}_{ik} G A \alpha_i \beta_i^* G,$$

where $d_i$'s are $O_<(1)$, $k$-independent generic constants and $\alpha_i, \beta_i$ are either $e_i$ or $h_i$. Due to similarity, we focus our discussion on the first inequality in (B.2) and briefly explain the other terms.

We start the argument of first inequality in (B.2). In view of (B.5), we find that every term in the first quantity in (B.2) has either one of the following two forms:

$$d_i \frac{1}{N} \sum_k h_{ik} e_k^* X_i G A \alpha_i \beta_i^* \tilde{B}^{(i)} R_i G e_i = \frac{d_i}{N} (\check{h}_i^* X_i G A \alpha_i) (\beta_i^* \tilde{B}^{(i)} R_i G e_i),$$

and

$$d_i \frac{1}{N} \sum_k h_{ik} e_k^* X_i G A R_i \tilde{B}^{(i)} \alpha_i \beta_i^* G e_i = \frac{d_i}{N} (\check{h}_i^* X_i G A R_i \tilde{B}^{(i)} \alpha_i) (\beta_i^* G e_i),$$

where used the definition of $\check{h}_i$ in (B.7). Then we provide controls for terms on the right-hand side of the above two equations. Since $\|X_i\|$ is bounded and $\|\check{h}_i\|_2 = 1$, by (B.3) and (v) of Assumption 2.4 we find that for some constant $C > 0$,

$$|\check{h}_i^* X_i G A \alpha_i| \leq \|X_i \check{h}_i\| \|G A \alpha_i\| \leq C (\alpha_i^* G^* G A \alpha_i)^{1/2},$$

$$|\beta_i^* \tilde{B}^{(i)} R_i G e_i| \leq C (\alpha_i^* G^* G e_i)^{1/2},$$

$$|\check{h}_i^* X_i G A R_i \tilde{B}^{(i)} \alpha_i| \leq C (\alpha_i^* R_i \tilde{B}^{(i)} G A B R_i \alpha_i)^{1/2},$$

$$|\beta_i^* G e_i| \leq C (\alpha_i^* G^* G e_i)^{1/2}.$$

It remains to study the bounds in (B.7) and (B.8). We first study the bounds of the first terms in (B.7) and (B.8). When $\alpha_i = e_i$, we have that for some constant $C > 0$

$$\alpha_i^* G^* G A \alpha_i = a_i e_i^* G^* A G e_i \leq C \frac{\text{Im} G_{ii}}{\eta},$$

where in the first step we used (B.3) and in the second step we used (v) of Assumption 2.4. Ward identity and (2.5). Similarly, we have that

$$\alpha_i^* R_i \tilde{B} A G^* G A B R_i \alpha_i = \check{h}_i^* G^* \tilde{B} A^2 \tilde{B} G h_i = e_i^* B^{1/2} \tilde{G} B^{1/2} \tilde{B} A^2 \tilde{B} B^{1/2} e_i \leq C \frac{\text{Im} G_{ii}}{\eta}.$$

Analogously, when $\alpha_i = h_i$, we have that

$$\alpha_i^* G A^* G A \alpha_i = e_i A^* \tilde{G} A e_i = e_i^* \tilde{A} B^{1/2} \tilde{G} B^{-1} \tilde{G} B^{1/2} \tilde{A} \tilde{B} \tilde{G} e_i = b_i^{-1} e_i^* \tilde{A} B \tilde{G} \tilde{B} \tilde{G} e_i \leq C \frac{\text{Im} G_{ii}}{\eta},$$

$$\alpha_i^* R_i \tilde{B} A G^* G A B R_i \alpha_i = e_i^* G^* G^* \tilde{B} A^2 \tilde{B} G e_i \leq C \frac{\text{Im} G_{ii}}{\eta}.$$

Moreover, for the second terms in (B.7) and (B.8), we readily see that

$$e_i^* G^* G e_i \leq C \frac{\text{Im} G_{ii}}{\eta}.$$

Combining all the above bounds, we conclude that

$$\frac{1}{N} \sum_{(i)} e_k^* X_i \Delta_G(i,k) e_i < \frac{1}{N} \frac{\text{Im} G_{ii} + \text{Im} G_{ii}}{\eta} = \Pi_i^2.$$
Then we briefly discuss the proof of the rest of the inequalities. In fact, for the remaining estimates, we will replace $\Delta_G$ as a sum of quantities in (B.6) and estimate every term of the summands. Specifically, for the second inequality, all terms are of either one of the following forms:

$$\frac{d_i}{N} \sum_k (\check{h}_{ik} e^*_i X_i G e_i)\text{tr}(DX_A G A \alpha_i \beta_i^* \check{B}^{(i)} R_i G) = \frac{d_i}{N^2} (\check{h}_{i} X_i G e_i)(\beta_i^* \check{B}^{(i)} R_i GDX_A G A \alpha_i),$$

$$\frac{d_i}{N} \sum_k (\check{h}_{ik} e^*_i X_i G e_i)\text{tr}(DX_A GAR_i \check{B}^{(i)} \alpha_i \beta_i^* G) = \frac{d_i}{N^2} (\check{h}_{i} X_i G e_i)(\beta_i^* GAR_i \check{B}^{(i)} \alpha_i).$$

Similarly, all terms of the first quantity in (B.3) can be written in the same form except that the factor $(\check{h}_{i} X_i G e_i)$ needs to be replaced by $g_{ik}$ or $b_{ik}$ respectively according to $X_i = I$ or $\check{B}^{(i)}$. As $\|DX_A G^* R_i \check{B}^{(i)} \beta_i\|$ and $\|DX_A G^* \beta_i\|$ are $O_\prec(\eta^{-1})$ terms, the second inequality of (B.2) follows from an argument similar to the first inequality in (B.2) using

$$|\check{h}_{i} X_i G e_i| < \|Ge_i\| = \frac{\text{Im} G_{ii}}{\eta}.$$  

Analogously, the first inequality in (B.3) follows from the same reasoning with the bound above with the fact $g_{ik} \prec N^{-1/2} \prec \text{Im} G_{ii}/\eta$. For the second inequality in (B.3), we get the following two types for all the terms

$$\frac{d_i}{N} (\check{h}_{i} X_i G e_i)(e_i^* X_A G A \alpha_i)(\beta_i^* \check{B}^{(i)} R_i G e_i),$$

$$\frac{d_i}{N} (\check{h}_{i} X_i G e_i)(e_i^* X_A GAR_i \check{B}^{(i)} \alpha_i)(\beta_i^* G e_i).$$

Similarly, for (B.4), all terms are given by

$$\frac{d_i}{N} (\check{h}_{i} X_i G e_i)(\check{h}_{i}^* G A \alpha_i)(\beta_i^* \check{B}^{(i)} R_i G e_i),$$

$$\frac{d_i}{N} (\check{h}_{i} X_i G e_i)(\check{h}_{i}^* GAR_i \check{B}^{(i)} \alpha_i)(\beta_i^* G e_i).$$

In all of these cases, the proof follows from an analogous discussion as for the previous inequalities and the estimate

$$\check{h}_{i} X_i G e_i = O_\prec(1),$$

which follows from Assumption 6.2 and the fact that the left-hand side of the above equation is either $\check{S}_i$ or $T_i$. The completes our proof.

**Lemma B.3.** Let $X_i$ be $\check{B}^{(i)}$ or $I$, $X_A$ be $A$ or $I$ or $A^{-1}$. Suppose that $D$ is a random diagonal matrix satisfying $\|D\| \prec 1$. Then under the assumptions of Lemma B.2, we have

$$\frac{1}{N} \sum_k (\frac{\|g_{ik}\|^{-1}}{\partial g_{ik}}) e^*_i X_i G e_i \prec \frac{1}{N},$$

$$\frac{1}{N} \sum_k e^*_i X_i G e_i \text{tr}(DX_A \frac{\partial G}{\partial g_{ik}}) \prec \Pi^2 \Psi^2,$$

$$\frac{1}{N} \sum_k e^*_i X_i G e_i \text{tr}(DX_A \frac{\partial G}{\partial g_{ik}}) \prec \Pi^2 \Psi^2,$$

$$\frac{1}{N} \sum_k e^*_i X_i G e_i \frac{\partial T_i}{\partial g_{ik}} \prec \Pi^2,$$
Proof. We start with the first inequality in (B.10). Using the identity

\[ \frac{\partial \|g_i\|^{-1}}{\partial g_{ik}} = -\frac{g_{ik}}{2 \|g_i\|^2}, \]

and the definitions in (6.7) and (6.13), we find that

\[
\frac{1}{N} \sum_k (i) \frac{\partial \|g_i\|^{-1}}{\partial g_{ik}} e_k^* X_i Ge_i = -\frac{1}{\|g_i\|^2} \frac{1}{2N} \sum_k g_{ik} e_k^* X_i Ge_i = -\frac{1}{2N \|g_i\|^2} h_i^* X_i Ge_i
\]

\[ \left\{ \begin{array}{l}
-\frac{1}{2N} S_i \quad \text{if } X_i = \bar{B}^{(i)}, \\
-\frac{1}{2N} T_i \quad \text{if } X_i = I.
\end{array} \right. \]

It suffices to control \( S_i \) and \( T_i \). By the definition of \( T_i \) in (6.13), \( T_i < 1 \) follows from Assumption 6.2. Moreover, under Assumption 6.2 by (6.28), (6.29), (6.41) and (6.46), we have

\[ S_i = -(\bar{B}G)_{ii} + G_{ii} + T_i + e_i = \left( \frac{\Omega_B}{z} - 1 \right) \frac{1}{a_i - \Omega_B} + T_i + O_{\prec}(N^{-\gamma/4}) < 1, \]

where we used (i) of Proposition 3.1. This proves the first inequality in (B.10).

For the second inequality in (B.10), by (6.39), we have that

\[
\frac{1}{N} \sum_k (i) e_k^* X_i Ge_i \text{tr}(DX_A \partial G, \partial g_{ik})
\]

\[ = \frac{\ell^2}{\|g_i\|} \frac{1}{N} \sum_k (i) e_k^* X_i Ge_i \text{tr} \left( DX_A GA(e_i + h_i)^* \bar{B}^{(i)} R_i + R_i \bar{B}^{(i)} e_k(e_i + h_i)^* G + DX_A \Delta_G \right). \]

In view of (B.13), by (6.41) and Lemma 3.2, the contribution of the last term with \( \Delta_G \) is \( O_{\prec}(\Pi^2 \Psi^2) \). Note that the first term can be written as

\[
\frac{1}{N} \sum_k (i) \text{tr}(DX_A GA e_k(e_i + h_i)^* \bar{B}^{(i)} R_i G) e_k^* X_i Ge_i = \frac{1}{N^2} \sum_k (i) (e_i + h_i)^* \bar{B}^{(i)} R_i DX_A GA e_k e_k^* X_i Ge_i
\]

\[ = \frac{1}{N^2} (e_i + h_i)^* \bar{B}^{(i)} R_i DX_A GA I_i X_i Ge_i = \frac{1}{N^2} (-b_i h_i - \bar{B} e_i)^* DX_A GA I_i X_i Ge_i, \]

where we denoted \( I_i = I - e_i e_i^* \) and in the last step we used (6.9). Using the elementary bound that

\[ \|DX_A GA I_i X_i\| \leq C \|G\| \leq \frac{C}{\eta}, \]

and by a discussion similar to (B.9), for some constant \( C > 0 \), we can bound

\[ \frac{1}{N^2} (-b_i h_i - \bar{B} e_i)^* GDX_A GA I_i X_i Ge_i \leq \frac{C}{N^2 \eta} \left( \|G^* h_i + G^* (\bar{B})^{1/2} e_i\| \right) \|Ge_i\| \]

\[ \leq C \frac{1}{N^2 \eta} (h_i^* G^* h_i + e_i^* (\bar{B})^{1/2} G^* (\bar{B})^{1/2} e_i + e_i^* G^* G e_i)
\]

\[ = C \frac{1}{N^2 \eta} (e_i^* B^{-1/2} \bar{\gamma} B G^* B^{-1/2} e_i + e_i^* A^{-1/2} \bar{G} \bar{H} G^* A^{-1/2} e_i + e_i^* A^{-1/2} \bar{G}^* A \bar{G} A^{-1/2} e_i)
\]

\[ \leq C \frac{1}{N^2 \eta^2} \text{Im}(\bar{g}_{ii} + \bar{G}_{ii}) = C \Pi^2 \Psi^2. \]
Similarly, we can bound the second term using the identity
\[
\frac{1}{N} \sum_{k}^{(i)} e_k^i X_i G e_i \text{tr}( DXA GAR_i \tilde{B}^{(i)} e_k (e_i + h_i)^* G ) = \frac{1}{N^2} (e_i + h_i)^* GDXA GAR_i \tilde{B}^{(i)} I_i X_i G e_i.
\]

Regarding the first inequality in (B.11), its proof is similar to the second inequality in (B.10) with minor modification. For instance, the counterpart of (B.14) is
\[
\frac{1}{N} \sum_{k}^{(i)} \text{tr}(DXA GAe_k (e_i + h_i)^* \tilde{B}^{(i)} R_i G)e_k^* X_i g_i = \frac{1}{N^2} (-b_i h_i - \tilde{B} e_i)^* GDXA GA I_i X_i g_i.
\]
Compared to (B.15), we can simply replace the factor \( \|G e_i\| \) by \( \|g_i\| \prec 1 \prec (\text{Im} G_{ii}/\eta)^{1/2} \) in the first step. For the second inequality in (B.11), we again use (B.9) and a discussion similar to (B.14) to get
\[
\frac{1}{N} \sum_{k}^{(i)} e_k^i X_i G e_i e_k^* X A \partial G \partial_{g_{ik} e_i} = (B.16)
\]
\[
\begin{align*}
&\quad = \frac{\ell^2}{\|g_i\|} \frac{1}{N} \sum_{k}^{(i)} e_k^i X_i G e_i e_k^* X A G (e_k (e_i + h_i)^* \tilde{B}^{(i)} R_i + \tilde{B}^{(i)} e_k (e_i + h_i)^* ) G e_i + O_\prec (\Pi_i^2) \\
&\quad = \frac{\ell^2}{\|g_i\|} \frac{1}{N} \sum_{k}^{(i)} e_k^i X A G I_i X_i G e_i (e_i + h_i)^* \tilde{B}^{(i)} R_i G e_i + e_k^i X A G R_i \tilde{B}^{(i)} I_i X_i G e_i (e_i + h_i)^* G e_i + O_\prec (\Pi_i^2),
\end{align*}
\]
where in the second step we used Lemma (B.2). To bound the above quantity, on one hand, by (6.9) and (6.13), under Assumption 6.2, we can finish the proof.

On the other hand, using the bound that
\[
\|G^* X_A e_i\| \|G e_i\| \leq (e_i G G^* e_i)^{1/2} (e_i G^* G e_i)^{1/2}
\]
\[
= \left( e_i^* A^{1/2} G A^{-1/2} G A^{1/2} e_i \right)^{1/2} \left( A^{-1/2} \tilde{G} A G A^{-1/2} e_i \right)^{1/2} \leq C \frac{\text{Im} G_{ii}}{\eta},
\]
for some constant \( C > 0 \), and the fact that both \( \|A I_{i} X_i\| \) and \( \|A R_i \tilde{B}^{(i)} I_i X_i\| \) are bounded, we conclude that for some constant \( C > 0 \)
\[
|e_k^i X A G A I_i X_i G e_i| \leq C \frac{\text{Im} G_{ii}}{\eta}, \quad e_k^i X A G R_i \tilde{B}^{(i)} I_i X_i G e_i \leq C \frac{\text{Im} G_{ii}}{\eta}.
\]
Togethr with (B.16) and (B.17), we can complete the proof.

For the inequality in (B.12), we note that
\[
\frac{\partial T_i}{\partial g_{ik}} = \frac{\partial h_i^*}{\partial g_{ik}} G e_i + h_i^* \frac{\partial G}{\partial g_{ik}} e_i.
\]
Due to similarity, we only study the summand containing the first term, i.e.,
\[
\begin{align*}
\frac{1}{N} \sum_{k}^{(i)} e_k^i X_i G e_i \frac{\partial h_i^*}{\partial g_{ik}} G e_i &= \frac{1}{N} \sum_{k}^{(i)} e_k^i X_i G e_i \frac{\partial g_{ik}}{\partial g_{ik}}^{-1} \partial h_i^* G e_i = \frac{1}{N} \sum_{k}^{(i)} \tilde{h}_i^* e_k^i X_i G e_i h_i^* G e_i \\
&= -\frac{1}{2\|g_i\|} \tilde{h}_i^* X_i G e_i h_i^* G e_i = \left\{ \begin{array}{ll} -\frac{1}{2\|g_i\| N} \tilde{S}_iT_i & \text{if } X_i = \tilde{B}^{(i)}; \\
-\frac{1}{2\|g_i\| N} \tilde{T}_iT_i & \text{if } X_i = I.
\end{array} \right.
\end{align*}
\]
where we used (A.3). Since \( \tilde{S}_i, \tilde{T}_i, T_i \) are all \( O_\prec (1) \) under Assumption 6.2, we can finish the proof.
Following the same proof strategy of Lemma B.3, we can prove the following result for the errors arising in the expansion of off-diagonal entries. We omit the details of the proof.

**Lemma B.4.** Let $X_i$ be $B^{(i)}$ or $I$, $X_A$ be $A$ or $I$ or $A^{-1}$, and $D = (d_{ik})$ be a random diagonal matrix with $\|D\| < 1$. Fix $z \in \mathcal{D}_\tau(\eta_L, \eta_U)$ and $j \neq i$, and assume that (6.23) and (6.30) hold. Then we have

\[
\begin{align*}
\frac{1}{N} \sum_{k} (i) \frac{\partial g_i}{\partial g_{ik}} e^*_k X_i Ge_j &< \frac{1}{N} , \\
\frac{1}{N} \sum_{k} e^*_k X_i Ge_j (\partial G / \partial g_{ik}) e_j &< (\Pi_i + \Pi_j)^2 , \\
\frac{1}{N} \sum_{k} e^*_k X_i Ge_j (\partial G / \partial g_{ik}) &< (\Pi_i + \Pi_j)^2 .
\end{align*}
\]

### B.2 Stability of perturbed linear system

In this section, we prove a stability result concerning the perturbations of the linear system \[\text{(3.1)}\] for a sufficiently large $\eta$, i.e. Lemma B.6. The stability result will serve as the starting point of our bootstrapping arguments in Section 3.3 for the system \[\text{(3.2)}\] and in Section 7.2 for \[\text{(7.30)}\].

The proof of Lemma B.6 relies on the well-known Kantorovich’s theorem. We record it in the following lemma, modified to our setting.

**Lemma B.5.** Let $\mathcal{C} \subseteq \mathbb{C} \times \mathbb{C}$ and $F : \mathcal{C} \to \mathbb{C} \times \mathbb{C}$ be a continuous function which is also continuously differentiable on $\text{int}(\mathcal{C})$, where $\text{int}(\mathcal{C})$ denotes the interior of $\mathcal{C}$. For $x_0 \in \text{int}(\mathcal{C})$, and suppose that there exist constants $b, L > 0$ such that for the matrix form $D$ of the differential operator the following hold:

(i). $DF(x_0)$ is non-singular;

(ii). $\|DF(x_0)^{-1}(DF(y) - DF(x))\| \leq L \|y - x\|$ for all $x, y \in \mathcal{C}$;

(iii). $\|DF(x_0)^{-1}F(x_0)\| \leq b$;

(iv). $2bL < 1$.

Denote

\[
t_\ast := \frac{1 - \sqrt{1 - 2bL}}{L} .
\]

If $\text{Ball}(x_0, t_\ast) \subseteq \mathcal{C}$, then there exists an unique $x_\ast \in \text{Ball}(x_0, t_\ast)$ such that $F(x_\ast) = 0$, where $\text{Ball}(x, t)$ denotes the ball in $\mathbb{C} \times \mathbb{C}$ of radius $t$ centered at $x$.

**Proof.** See Theorem 1 of [25] and the reference therein.

Then we state the main result of this subsection.

**Lemma B.6.** For $\eta > 0$ and $\theta \in (0, \pi/2)$, define

\[
\mathcal{E}(\eta, \theta) := \{ z \in \mathbb{C}_+ : \text{Im} z > \eta, \theta < \arg z < \pi - \theta \} .
\]

Let $(\mu_1, \mu_2)$ be either $(\mu_A, \mu_B)$ or $(\mu_A, \mu_B)$, and let $\Phi$ be either $\Phi_{AB}$ or $\Phi_{AB}$ accordingly. Moreover, we set $\Omega_1$ and $\Omega_2$ be analytic functions mapping $\mathcal{E}(\eta_1, \theta)$, for some $\eta_1$ and $\theta$, to $\mathbb{C}_+$. Denote $\tilde{\tau}(z) := \Phi(\Omega_1(z), \Omega_2(z), z)$.

Assume that there exists a positive constant $0 < c < 1$ such that the following hold for all $z \in \mathcal{E}(\eta_1, \theta)$:

\[
\left| \frac{\tilde{\Omega}_1(z)}{z} - 1 \right| \leq c , \quad \left| \frac{\tilde{\Omega}_2(z)}{z} - 1 \right| \leq c , \quad ||\tilde{\tau}(z)|| \leq c .
\]
Then there exist \( \eta_1 > \tilde{\eta}_1 \) and \( \theta > \tilde{\theta} \) depending only on \( \mu_\alpha, \mu_\beta \) and \( c \) such that for all sufficiently large \( N \), we have
\[
|\tilde{\Omega}_1(z) - \Omega_1(z)| \leq 2 \|f(z)\|, \quad |\tilde{\Omega}_2(z) - \Omega_2(z)| \leq 2 \|f(z)\|, \quad (B.21)
\]
hold for all \( z \in \mathcal{E}(\eta_1, \theta) \), where \( \Omega_1 \) and \( \Omega_2 \) are subordination functions corresponding to the pair \( (\mu_1, \mu_2) \) via Lemma 2.6.

Proof. Let \( \eta_1 > \tilde{\eta}_1 \) be some sufficiently large constant. For each \( z \) with \( \text{Im} \ z > \eta_1 \), denote \( \Phi(z, \cdot) := \Phi(\cdot, z) \). Our proof can be divided into two steps. In the first step, we apply Lemma B.5 to the function \( \Phi \) with initial value \( x_0 = (\tilde{\Omega}_1(z), \tilde{\Omega}_2(z)) \) to conclude that there exists an unique solution \( x_* \) on a properly chosen open set around \( x_0 \). In the second step, we prove that the obtained \( x_* \) coincides with \( x_* = (\Omega_1(z), \Omega_2(z)) \).

**Step 1:** In this step, we will justify that conditions (ii)-(iv) of Lemma B.5 hold on a properly chosen open set \( C_z \subset \mathbb{C}_+ \times \mathbb{C}_+ \) (c.f. (B.23) and \( \text{Ball}(x_0, t_4) \subset \mathbb{C}_+ \)). Note that (i) automatically holds once we have checked (ii) and (iii). Specifically, we will establish upper bounds of the following two quantities on \( C_z \) (i.e. conditions (ii) and (iii))
\[
\left\| D\Phi_z(\tilde{\Omega}_1(z), \tilde{\Omega}_2(z))^{-1}D\Phi_z(\tilde{\Omega}_1(z), \tilde{\Omega}_2(z)) \right\|, \quad \left\| D\Phi_z(\tilde{\Omega}_1(z), \tilde{\Omega}_2(z))^{-1} \right\| \sup_{(\omega_1, \omega_2) \in C_z} \left\| D^2\Phi_z(\omega_1, \omega_2) \right\|, \quad (B.22)
\]
and show that their product is bounded by \( 1/2 \) (i.e. condition (iv)).

We start verifying condition (iii) of Lemma B.5. By (B.24) to Lemma B.5 we have that
\[
\left\| D\Phi_z(\tilde{\Omega}_1(z), \tilde{\Omega}_2(z))^{-1} \right\| = \frac{1}{|1 - z^2 \mathcal{L}_{\mu_1}(\tilde{\Omega}_2(z))\mathcal{L}_{\mu_2}(\tilde{\Omega}_1(z))|} \left\| \begin{pmatrix} 1 & -z \mathcal{L}_{\mu_1}(\tilde{\Omega}_2(z)) \\ \mathcal{L}_{\mu_2}(\tilde{\Omega}_1(z)) & 1 \end{pmatrix} \right\|. \quad (B.23)
\]
Moreover, under (v) of Assumption 2.3 by Lemmas 3.4 and 3.11 we have that
\[
|z \mathcal{L}_{\mu_1}(\tilde{\Omega}_2(z))| = \left| \frac{\tilde{\Omega}_2(z)}{|\tilde{\Omega}_2(z)|} \int (x - \tilde{\Omega}_2(z))^2 d\mu_1(x) \right| \leq \frac{z}{|\tilde{\Omega}_2(z)|} \left( \int \left[ \frac{x}{|x - \tilde{\Omega}_2(z)|} + \frac{1}{|x - \tilde{\Omega}_2(z)|} \right] d\mu_1(x) \right) \leq \frac{1}{1 - c} \left( \frac{E^\alpha_x + \delta}{|\tilde{\Omega}_2(z)| - (E^\alpha_x + \delta)^2} + \frac{1}{|\tilde{\Omega}_2(z)| - (E^\alpha_x + \delta)} \right) \hat{\mu}_1(\mathbb{R}_+),
\]
where in the third step we used (B.20). Using the assumption \( |\tilde{\Omega}_2(z)/z| > 1 - c \) and the fact \( \eta_1 \) is sufficiently large, we have that for some constant \( C > 0 \),
\[
|z \mathcal{L}_{\mu_1}(\tilde{\Omega}_2(z))| \leq C |z|^{-1} \leq C \eta_1^{-1},
\]
holds for all \( z \in \mathcal{E}(\eta_1, \tilde{\theta}) \). Together with (B.23), we conclude that for some constant \( C > 0 \),
\[
\left\| D\Phi_z(\tilde{\Omega}_1(z), \tilde{\Omega}_2(z))^{-1} \right\| \leq (1 - C \eta_1^{-2})^{-1} (2 + 2C \eta_1^{-2})^{1/2} \leq 2,
\]
where we simply bounded the operator norm using its Frobenius norm in the first step and used the assumption that \( \eta_1 \) is sufficiently large in the second step. Thus we have established that
\[
\left\| D\Phi_z(\tilde{\Omega}_1(z), \tilde{\Omega}_2(z))^{-1} \Phi_z(\tilde{\Omega}_1(z), \tilde{\Omega}_2(z)) \right\| \leq 2 \|f(z)\| =: b. \quad (B.24)
\]

Next, we justify conditions (ii) and (iv) of Lemma B.5. Recall (B.24). We see that
\[
\left\| D^2\Phi_z(\omega_1, \omega_2) \right\| = \left\| D z \mathcal{L}_{\mu_1}(\tilde{\Omega}_2(z)), z \mathcal{L}_{\mu_2}(\tilde{\Omega}_2(z)) \right\| = \left\| \begin{pmatrix} z \mathcal{L}_{\mu_1}(\omega_2) & 0 \\ 0 & z \mathcal{L}_{\mu_2}(\omega_1) \end{pmatrix} \right\|.
\]
By Lemma 4.3 and (v) of Assumption 2.4, we have
\[ |L''_{\mu_1}(\omega_1)| \leq \int_{\mathbb{R}^+} \frac{1}{|x - \omega_1|^3} \delta \omega_1(x) \leq C \frac{1}{|\omega_1|} - E^\alpha \delta \].

Based on the above discussion, we see that for all \( z \in \mathcal{E}(\eta_1, \tilde{\theta}) \), the function \( \Phi(z) \) is \( C_0 |z|^{-2} \)-Lipschitz in the domain
\[ C_z := \{ (\omega_1, \omega_2) \in \mathbb{C}^2 : \frac{|\omega_1|}{|z|} \geq 1 - \frac{c}{2}, \frac{|\omega_2|}{|z|} \geq 1 - \frac{c}{2} \} , \]
for some constant \( C_0 \) depending only on \( \mu_\alpha, \mu_\beta \), and \( c \). Therefore, we have that
\[ \| \Phi(z_1) \| \mid \Phi(z_2) \| \leq 2 C_0 |z|^{-2} \leq 2 C_0 \eta_1^{-2} =: L. \]

Moreover, under the assumption (B.20), it is clear that \( (\tilde{\Omega}_1(z), \tilde{\Omega}_2(z)) \in C_z \) for all \( z \in \mathcal{E}(\eta_1, \tilde{\theta}) \), and \( b \) and \( L \) defined in (B.24) and (B.26) can be bounded by
\[ Lb \leq 4 C_0 \eta_1^{-2} \| \Phi(z) \| \leq 4 C_0 c \eta_1^{-2}. \]

Since \( \eta_1 \) is sufficiently large, we conclude that condition (iv) of Lemma 4.5 holds for our choices of \( b \) and \( L \).

To complete the argument of Step 1, we need to show that the closure of \( B((\tilde{\Omega}_1(z), \tilde{\Omega}_2(z)), t_\ast) \) is contained in \( C_z \). Substituting \( b = 2 \| \Phi(z) \| \) and \( L = 2 C_0 \eta_1^{-2} \) in (B.18), we get
\[ t_\ast = \frac{1 - \sqrt{1 - 4 C_0 \eta_1^{-2} \| \Phi(z) \|}}{2 C_0 \eta_1^{-2}} \leq 2 \| \Phi(z) \| \leq 2 c, \]
where we used the assumption that \( \eta_1 \) is sufficiently large and (B.20). We now choose \( \theta > \tilde{\theta} \) such that \( \theta \) is sufficiently close to \( \pi/2 \). Since \( |z| = \csc \theta \text{ Im } z \) for \( z \in \mathcal{E}(\eta_1, \theta) \), by (B.20), we have
\[ \text{Im } \tilde{\Omega}_1(z) = \text{Im } z + \text{Im } z \left( \frac{\tilde{\Omega}_1(z)}{z} - 1 \right) \geq \text{Im } z - c|z| \geq \eta_1(1 - c \csc \theta). \]

Consequently, the condition \( |\omega_1 - \tilde{\Omega}_1(z)| \leq t_\ast \) implies that
\[ \text{Im } \omega_1 \geq \text{Im } \tilde{\Omega}_1(z) - 2 \| \Phi(z) \| \geq \eta_1(1 - c \csc \theta) - 2c > 0, \]
for all \( z \in \mathcal{E}(\eta_1, \theta) \). This shows that \( \omega_1 \in \mathbb{C}^+ \). On the other hand, \( |\omega_1 - \tilde{\Omega}_1(z)| \leq t_\ast \) also yields that
\[ \frac{|\omega_1(z)|}{|z|} \geq 1 - c - \frac{2c}{|z|} \geq 1 - c - \frac{2c}{\eta_1} > 1 - \frac{c}{2}, \]
where we used (B.28). Similar results hold for \( \omega_2 \) satisfying \( |\omega_2 - \tilde{\Omega}_2(z)| \leq t_\ast \). This proves that
\[ \{ (\omega_1, \omega_2) \in \mathbb{C}^2 : \| (\tilde{\Omega}_1(z), \tilde{\Omega}_2(z)) - (\omega_1, \omega_2) \| \leq t_\ast \} \subset C_z. \]

This finishes that proof of Step 1.

**Step 2.** In step 2, we have shown that the conditions of Lemma 4.5 are satisfied. Therefore, we conclude that there exists a solution \( (\omega_1(z), \omega_2(z)) \) for the equation \( \Phi(\cdot, \cdot, z) = 0 \) such that
\[ \| (\omega_1(z), \omega_2(z)) - (\tilde{\Omega}_1(z), \tilde{\Omega}_2(z)) \| \leq t_\ast. \]
In this step, we prove that \((\omega_1(z), \omega_2(z))\) coincides with \((\Omega_1(z), \Omega_2(z))\). As \(\Phi(\omega_1(z), \omega_2(z), z) = 0\), we find that \(\omega_1(z)/z\) satisfies that

\[
\frac{\omega_1(z)}{z} = L_{\mu_1}(\omega_2(z)) = L_{\mu_1}(zL_{\mu_2}(\omega_1(z))).
\]

That is to say, \(\omega_1(z)/z\) is a fixed point of the function

\[
\omega \mapsto F_z(\omega) := L_{\mu_1}(zL_{\mu_2}(z\omega)).
\]

By Lemmas 3.5 and 3.11, the definition of \(L_{\mu}(z)\) in (2.4) and the facts that \(L_{\mu}(z) \in \mathbb{C}_+\) and \(M_{\mu}(z) \in \mathbb{C}_+\), we have that

\[
0 < \arg L_{\mu}(z) < \pi - \arg z, \quad z \in \mathbb{C}_+.
\]

Based on the above observation, we find that the function \(F_z\) maps the sector \(\{\omega \in \mathbb{C}_+ : 0 < \arg \omega < \pi - \arg z\}\) to itself. It is clear that the sector is conformally equivalent to the unit disk. Since \(F_z\) is a non-constant analytic map, by Schwarz-Pick lemma, \(F_z\) can have at most one fixed point in the sector. Moreover, on one hand, as \(\Phi(\omega_1(z), \omega_2(z), z) = 0\), we have

\[
\text{Im} \frac{\omega_1(z)}{z} = \text{Im} L_{\mu_1}(\omega_2(z)) > 0,
\]

which implies that \(\omega_1(z)/z\) is in the sector; on the other hand, \(\Omega_1(z)/z\) is already a fixed point in the sector. This proves that \(\omega_1(z) = \Omega_1(z)\). Similarly, we can show that \(\omega_2(z) = \Omega_2(z)\). This finishes the proof of Step 2. \(\square\)

### C Bootstrapping continuity argument

#### C.1 Proof of Lemmas 8.3, 8.4 and 8.5

In this subsection, we prove Lemmas 8.3, 8.4 and 8.5. We begin with Lemma 8.3. The proof strategy is the same with Lemma 8.3 of [1], which needs three technical inputs. The first input is Lemma C.1 which is a weaker analog of Proposition 7.4 and will be used in the final stage of the proof of Lemma 8.3. The other two ingredients are Lemmas C.2 and C.3 below. They are cutoff versions of Proposition 6.3 and Proposition 7.2 respectively; see the cutoff function in (C.11). We first state these results and prove Lemma 8.3 in this subsection. The proofs of Lemmas C.1, C.2 and C.3 are offered in Section C.2.

**Lemma C.1.** Let \(z \in D_{r(\eta, \eta)}, \epsilon \in (0, \gamma/12),\) and \(k \in (0, 1]\) be fixed values. Let \(\hat{\Lambda}(z)\) be some deterministic positive control parameter such that \(\hat{\Lambda}(z) \leq N^{-\gamma/4}\). Suppose that \(\Lambda(z) \leq \hat{\Lambda}(z)\) and

\[
|S_{AB}A_t + f_tA_t^2 + O(\Lambda_t^3)| \leq N^{2\epsilon/5} \frac{|S_{AB}| + \hat{\Lambda}}{(N\eta)^k}, \quad t = A, B,
\]

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

(i) If \(\sqrt{k + \eta} > N^{-\epsilon}\hat{\Lambda}\), there is a sufficiently large constant \(K_0 > 0\) which is independent of \(z\) and \(N\), such that

\[
I \left( \Lambda \leq \frac{|S_{AB}|}{K_0} \right) |A_t| \leq C \left( N^{-2\epsilon}\hat{\Lambda} + \frac{N^{7\epsilon/5}}{(N\eta)^k} \right),
\]

holds on \(\hat{E}(z)\).

(ii) If \(\sqrt{k + \eta} \leq N^{-\epsilon}\hat{\Lambda}\), then we have that

\[
|A_t| \leq C \left( N^{-\epsilon}\hat{\Lambda} + \frac{N^{7\epsilon/5}}{(N\eta)^k} \right), \quad t = A, B.
\]

holds on \(\hat{E}(z)\).
Lemma C.2. Let small $\epsilon \in (0, \gamma/12)$ and large $D > 0$ be fixed. Recall (3.14). There exists an $N_{01}(\epsilon, D) \in \mathbb{N}$ such that the following hold for all $N \geq N_{01}(\epsilon, D)$: for all $z \in D_\tau(\eta_L, \eta_U)$, there exists an event $\Xi_1(z)$ with $\mathbb{P}[\Xi_1(z)] > 1 - N^{-D}$ so that the following hold on $\Xi_1(z) \cap \Theta\left(z, \frac{N_{3c}}{(N\eta)^{1/3}}, \frac{N_{3c}}{(N\eta)^{1/3}}\right)$:

$$\Lambda^c_0(z) \leq N^{c/2}\Psi(z), \quad \Lambda_T \leq N^{c/2}\Psi(z), \quad \Lambda^c_1 \leq N^{c/2}\Psi(z), \quad \Lambda_T \leq N^{c/2}\Psi(z), \quad \Upsilon(z) \leq N^{c/2}\Psi(z).$$

Denote the control parameters

$$\hat{\Lambda} := \frac{N^{3c}}{(N\eta)^{1/3}}, \quad \hat{\Pi} := \left(\frac{1}{(N\eta)^{2/3}} \left(\frac{\text{Im} m_{A, B}(z)}{(N\eta)^{-1/3}} \left(\frac{\Lambda}{(N\eta)^{-1/3}}\right) + N^{3c} \right) + \frac{1}{(N\eta)^{2/3}}\right)^{1/2}.$$ (C.4)

Lemma C.3. Let small $\epsilon \in (0, \gamma/12)$ and large $D > 0$ be fixed. Recall (3.35). There exists an $N_{02}(\epsilon, D) \in \mathbb{N}$ such that the following hold for all $N \geq N_{02}(\epsilon, D)$: for all $z \in D_\tau(\eta_L, \eta_U)$, there exists an event $\Xi_2(z)$ with $\mathbb{P}[\Xi_2(z)] > 1 - N^{-D}$ so that the following hold on $\Xi_2(z) \cap \Theta\left(z, \frac{N_{3c}}{(N\eta)^{1/3}}, \frac{N_{3c}}{(N\eta)^{1/3}}\right)$:

$$\Phi^c_\Lambda \leq N^{c/3}\hat{\Pi}, \quad \Phi^c_\Lambda \leq N^{c/3}\hat{\Pi}, \quad |\tilde{3}_1| \leq N^{c/3}\hat{\Pi}, \quad |\tilde{3}_2| \leq N^{c/3}\hat{\Pi}.$$

Armed with Lemmas C.1, C.2, and C.3 we proceed to the proof of Lemma 8.3.

Proof of Lemma 8.3 By (ii) and (iii) of Proposition 3.1 we have that $\text{Im} m_{A, B}(z) \sim |S_{AB}|$. For the parameters in (C.4), we have that for some constant $C > 0$

$$N^{c/3}\hat{\Pi} \leq CN^{c/3}\frac{|S_{AB}| + \hat{\Lambda}}{N\eta} + N^{c/3}\frac{|S_{AB}| + \hat{\Lambda}}{N\eta} = CN^{c/3}\frac{|S_{AB}| + \hat{\Lambda}}{N\eta} \leq CN^{c/3}\frac{|S_{AB}| + \hat{\Lambda}}{(N\eta)^{-1/3}},$$

where in the last step we used that $\sqrt{xy} \leq x + y, \; x, y \geq 0$. By a discussion similar to (7.33), we obtain that

$$\left|\frac{S_{A}}{z}A + \mathcal{T}_A\Lambda^3 + O(|\Lambda^3|)\right| \leq |\tilde{3}_1| + O(|\Phi^c_\Lambda|) + O(|\Phi^c_\Lambda|) \leq CN^{c/3}\hat{\Pi} \leq N^{2c/5}\frac{|S_{A}| + \hat{\Lambda}}{(N\eta)^{-1/3}},$$

on the event $\Theta\left(z, \frac{N^{3c}}{(N\eta)^{1/3}}, \frac{N^{3c}}{(N\eta)^{1/3}}\right)$, where we used (3.15) and (3.33), and the fact $\Lambda \leq N^{3c}/(N\eta)^{1/3}$.

We now start the proof of Lemma 8.3. In what follows, we will show that we can choose $\Xi(z, \epsilon) = \Xi_1(z) \cap \Xi_2(z)$ and $N_1 = \max\{N_{01}(\epsilon, D), N_{02}(\epsilon, D)\}$, where $\Xi_1(z)$ and $\Xi_2(z)$ are events from Lemma C.2 and Lemma C.3 respectively. Due to similarity, we only prove part (i). Denote

$$\Xi(z) = \begin{cases} \Xi_1(z) \cap \Xi_2(z) \cap \Theta\left(z, \frac{N^{3c}}{(N\eta)^{1/3}}, \frac{N^{3c}}{(N\eta)^{1/3}}\right) & \text{if } z \in D_>, \\ \Xi_1(z) \cap \Xi_2(z) \cap \Theta\left(z, \frac{N^{3c}}{(N\eta)^{1/3}}, \frac{N^{3c}}{(N\eta)^{1/3}}\right) & \text{if } z \in D_\leq. \end{cases}$$

By (C.2), for $k = 1/3$, $\hat{\Lambda} = N^{3c}/(N\eta)^{1/3}$, when $\sqrt{\kappa + \eta} > N^{-\varepsilon}\hat{\Lambda}$, i.e. $z \in D_>$, we have that for some constant $C > 0$

$$|\Lambda| \leq C \left(N^{-2\varepsilon}\hat{\Lambda} + \frac{N^{7\varepsilon/5}}{(N\eta)^{1/3}}\right),$$

holds on the event $\Xi(z)$, where we can ignore the indicator function since $\Lambda \leq N^{-\varepsilon/10}|S_{AB}|$. Thus we have that for some constant $C > 0$

$$|\Lambda| \leq CN^{-2\varepsilon}\hat{\Lambda} + \frac{N^{7\varepsilon/5}}{(N\eta)^{1/3}} \leq CN^{-8\varepsilon/5}\hat{\Lambda} \leq CN^{-3\varepsilon/5}\sqrt{\kappa + \eta}.$$
Consequently, we have that
\[
\lambda \leq C \min \left( \frac{N^{12/5}}{(N\eta)^{1/3}}, \frac{N^{-3\epsilon/5} \sqrt{\kappa + \eta}}{N^{1/3}} \right) \leq \min \left( \frac{N^{5\epsilon/2}}{(N\eta)^{1/3}}, \frac{N^{-\epsilon/2} |S|}{N^{1/3}} \right). \tag{C.5}
\]

Moreover, by Lemma C.2 (i) of Proposition 8.1 and (C.5), we see that
\[
\lambda_d(z) \leq \lambda_d^\alpha(z) + \frac{\lambda}{|a_i - \Omega_B(z) - \lambda|^2} \leq \frac{N^{5\epsilon/2}}{(N\eta)^{1/3}}, \quad \lambda_T(z) \leq N^{\epsilon/2} \Psi \leq \frac{N^{5\epsilon/2}}{\sqrt{N\eta}}. \tag{C.6}
\]

Similarly, we can show that
\[
\lambda_d(z) \leq \frac{N^{5\epsilon/2}}{(N\eta)^{1/3}}, \quad \lambda_T(z) \leq \frac{N^{5\epsilon/2}}{\sqrt{N\eta}}. \tag{C.7}
\]

Therefore, by (C.5), (C.6) and (C.7), using the definition (8.5), we have proved that
\[
\Xi(z) = \Xi_1(z) \cap \Xi_2(z) \cap \Theta > \left( z, \frac{N^{3\epsilon}}{(N\eta)^{1/3}}, \frac{N^{3\epsilon}}{\sqrt{N\eta}}, \frac{\epsilon}{10} \right) \subset \Theta > \left( z, \frac{N^{5\epsilon/2}}{(N\eta)^{1/3}}, \frac{N^{5\epsilon/2}}{\sqrt{N\eta}}, \frac{\epsilon}{2} \right).
\]

This completes the proof. \(\square\)

Then we prove Lemma 8.4.

**Proof of Lemma 8.4.** From Proposition 8.1 we find that Assumption 6.2 holds uniformly in \(z \in \mathcal{D}_r(\eta_L, \eta_U)\). Consequently, we find that Proposition 7.4 holds uniformly in \(z \in \mathcal{D}_r(\eta_L, \eta_U)\). More specifically, let \(\hat{\Lambda}(z)\) be a deterministic positive function with \(\Lambda(\bar{z}) < \hat{\Lambda}(z) \leq N^{-\gamma/4}\), we have that for \(\epsilon = A, B,\)
\[
\left| \frac{S_{AB}(z)}{z} \Lambda_i + T_i \Lambda_i(z)^2 + O(|\Lambda_i|^3) \right| < \Psi^2 \left( \sqrt{\operatorname{Im} m_{\mu, \delta_{\mu B}}(z) + \hat{\Lambda}(z)} \cdot |S_{AB}(z)| \Lambda(z) + \hat{\Lambda}(z) + \Psi^2 \right). \tag{C.8}
\]

holds uniformly in \(z \in \mathcal{D}_r(\eta_L, \eta_U)\). Moreover, since \(|S_{AB}(z)| \sim \sqrt{\kappa + \eta} \sim \operatorname{Im} m_{\mu, \delta_{\mu B}}(z)\) and \(\hat{\Lambda}(z) \geq (N\eta)^{-1}\), by (C.8), we have that
\[
\left| \frac{S_{AB}(z)}{z} \Lambda_i + T_i \Lambda_i(z)^2 + O(|\Lambda_i|^3) \right| < \frac{1}{N\eta} \left( |S_{AB}(z)| \Lambda(z) + \hat{\Lambda}(z) \right). \tag{C.9}
\]

Therefore, for fixed small \(\epsilon_0 \in (0, \gamma/12)\) and large \(D > 0\), we can find an \(N_0(\epsilon_0, D) \in \mathbb{N}\) such that for each \(z \in \mathcal{D}_r(\eta_L, \eta_U)\) and \(N \geq N_0(\epsilon_0, D),\) there exists an event \(\Xi(z, \epsilon_0, D)\) with \(\mathbb{P}[\Xi(z, \epsilon_0, D)] > 1 - N^{-D}\) on which we have
\[
\left| \frac{S_{AB}(z)}{z} \Lambda_i + T_i \Lambda_i(z)^2 + O(|\Lambda_i|^3) \right| \leq N^{\epsilon_0/3} \frac{1}{N\eta} \left( |S_{AB}(z)| \Lambda(z) + \hat{\Lambda}(z) \right) \leq N^{2\epsilon_0/5} |S_{AB}| + \frac{\hat{\Lambda}}{N\eta}. \tag{C.9}
\]

Using (C.8), by Lemma C.1 with \(k = 1\), we have shown that the results hold on \(\Xi(z, \epsilon_0, D)\). This completes our proof. \(\square\)

**Proof of Lemma 8.5.** The proof is similar to that of Lemma 8.4 except that we employ Lemma 8.4 as a technical input instead of Lemma C.1. We omit the details here. \(\square\)
C.2 Proof of Lemmas C.1, C.2 and C.3

We first prove Lemma C.1 in this subsection.

Proof of Lemma C.1 Notice that (C.1) together with the assumption \( \hat{\Lambda}(z) \leq N^{-\gamma/4} \) implies

\[
\left| \frac{S_{AB}(z)}{z} \Lambda_i + T_i \Lambda_i^2 \right| \leq N^{2/5} \frac{|S_{AB}(z)| + \hat{\Lambda}}{(N\eta)^k} + N^{-\gamma/4}\hat{\Lambda}^2, \quad i = A, B. \tag{C.10}
\]

We first prove (C.2). We choose \( K_0 \) such that

\[ K_0 \geq 2 \sup \{|zT_i(z)| : z \in \mathcal{D}_\tau(\eta_L, \eta_U)\}, \quad i = A, B. \]

By (iii) of Proposition 3.1, we find that \( K_0 \) is indeed finite. With such a choice of \( K_0 \), under the assumptions of (i), we find that

\[
\mathbb{I}\left( A \leq \frac{|S_{AB}|}{K_0} \right) |T_i \Lambda_i^2| \leq \frac{|S_{AB}|}{z} \Lambda_i \left| \frac{|zT_i|}{K_0} \right| \leq \frac{1}{2} \frac{|S_{AB}|}{z} \Lambda_i,
\]

where we used \( \Lambda_i \leq |S_{AB}|/K_0 \) and the definition of \( K_0 \) for the second and third step, respectively. This shows that the left-hand side of (C.10) is dominated by \( S_{AB} \Lambda_i / z \). Consequently, we have that

\[
\mathbb{I}\left( A \leq \frac{|S_{AB}|}{K_0} \right) \left| \frac{S_{AB}}{z} \Lambda_i \right| \leq 2 \left( N^{2/5} \frac{|S_{AB}| + \hat{\Lambda}}{(N\eta)^k} + N^{-\gamma/4}\hat{\Lambda}^2 \right).
\]

This further implies that for some constant \( C > 0 \)

\[
\mathbb{I}\left( A \leq \frac{|S_{AB}|}{K_0} \right) \left| \Lambda_i \right| \leq C \frac{1}{|S_{AB}|} \left( N^{2/5} \frac{|S| + \hat{\Lambda}}{(N\eta)^k} + N^{-\gamma/4}\hat{\Lambda}^2 \right) \leq C \left( \frac{N^{7/5}}{(N\eta)^k} + N^{\epsilon/4}\hat{\Lambda} \right),
\]

where we used \( \sqrt{k + \eta} > N^{-\epsilon}\hat{\Lambda} \) from the assumption and \( \sqrt{k + \eta} \sim |S| \) from (iii) of Proposition 3.1. Since \( \epsilon < \gamma/12 \), we conclude the proof of part (i).

We next handle (C.3). Denote

\[ X = -S_{AB} \frac{1}{z} \Lambda_i + T_i \Lambda_i^2. \]

Using the fact that \( |T_i| \geq c \) for some small constant \( c > 0 \), i.e., (iii) of Proposition 3.1 we can consider \( \Lambda_i \) as the solution of a quadratic equation involving \( X \) so that

\[
\Lambda_i = \frac{1}{2T_i} \left( -\frac{S_{AB}}{z} + \sqrt{\frac{S_{AB}^2}{z^2} - 4T_iX} \right),
\]

for an appropriate choice of branch cut. We thus have that for some constant \( C > 0 \),

\[
|\Lambda_i| \leq C \left( |S_{AB}| + \sqrt{|X|} \right) \leq C \left( \sqrt{k + \eta} + \sqrt{N^{2/5} \frac{|S_{AB}| + \hat{\Lambda}}{(N\eta)^k} + N^{-\gamma/4}\hat{\Lambda}^2} \right)
\]

\[
\leq C \left( N^{-\epsilon}\hat{\Lambda} + \sqrt{\frac{N^{2/5}\hat{\Lambda}}{(N\eta)^k} + N^{-\gamma/4}\hat{\Lambda}^2} \right) \leq C \left( N^{-\epsilon}\hat{\Lambda} + \frac{N^{7/5}}{(N\eta)^k} + N^{-\gamma/2}\hat{\Lambda} \right),
\]

where we used (C.10) in the second step, \( |S| \sim \sqrt{k + \eta} \leq N^{-\epsilon}\hat{\Lambda} \) in the third step, and \( \sqrt{x+y} \leq x+y \) and \( \sqrt{x^2+y^2} \leq \sqrt{x} + \sqrt{y} \) for \( x, y > 0 \) in the last step. Using \( \epsilon < \gamma/12 \), we complete the proof of part (ii). \( \square \)
Next, we prove Lemma C.2 which essentially generalizes Proposition 6.3 and its key input Lemma C.3. Our goal here is to replace the Assumption 6.2 with the event \( \Theta \left( z, \frac{N^{2\epsilon}}{(N_\eta)^{2\epsilon}}, \frac{N^{2\epsilon}}{\sqrt{N_\eta}} \right) \) which holds with high probability. We will follow the proof idea as shown in Lemma 8.3. In detail, if \( X \) is the quantity for which we want to establish a bound when restricted on \( \Theta \), we will bound the high order moments of \( I(\Theta_0)X \) instead of \( X \) directly, where \( \Theta_0 \) is an event such that \( \Theta \left( z, \frac{N^{2\epsilon}}{(N_\eta)^{2\epsilon}}, \frac{N^{2\epsilon}}{\sqrt{N_\eta}} \right) \subseteq \Theta_0 \). Then on the intersection of \( \Theta \) and \( \Theta_0 \) and some high-probability event on which the bound for \( I(\Theta_0)X \) holds, we can establish the bound for \( X \).

As we have seen in the proof of Proposition 6.3, we need to apply recursive moment estimate where differentiability is required. To seek for a smooth approximation for the indicator function, we introduce the following cutoff function. For a (large) constant \( K \) and some constant \( C > 0 \), we choose a smooth function \( \varphi : \mathbb{R} \to \mathbb{R} \) that satisfies for all \( x \in \mathbb{R} \)

\[
\varphi(x) = \begin{cases} 
0, & \text{if } |x| > 2K \\
1, & \text{if } |x| \leq K \, , \quad |\varphi'(x)| \leq CK^{-1}.
\end{cases}
\] (C.11)

Denote

\[
\Gamma_i = |G_{ii}| + |G_{iii}| + |T_i| + |\bar{T}_i|^2 + |\text{tr} G|^2 + |\text{tr}(GA)|^2 + |\text{tr}(\bar{B}G)|^2 ,
\] (C.12)

where \( \bar{T}_i \) is defined as an analogue of \( T_i \) (c.f. 6.13) by switching the roles of \( A \) and \( B \) and also \( U \) and \( U^* \).

**Proof of Lemma C.2** Recall (6.13) and (6.12). Denote

\[
\tilde{X}_i^{(p,q)} := \varphi(\Gamma_i)^{p-q} P_i \tilde{\Gamma}^i, \quad \tilde{Y}_i^{(p,q)} := \varphi(\Gamma_i)^{p+q} R_i^p \tilde{\Gamma}^i .
\]

We first claim that the following recursive moment estimates hold true for \( N \geq N_1(\epsilon_1, D_1) \), where \( N_1(\epsilon_1, D_1) \) is some large positive integer and \( \epsilon_1, D_1 > 0 \) will be chosen later in terms of \( \epsilon, D \),

\[
\begin{align*}
\mathbb{E} \left[ \tilde{X}_i^{(p,p)} \right] & \leq \mathbb{E} \left[ a_{i1} \tilde{X}_i^{(p-1,p)} \right] + \mathbb{E} \left[ a_{i2} \tilde{X}_i^{(p-2,p)} \right] + \mathbb{E} \left[ a_{i3} \tilde{X}_i^{(p-1,p-1)} \right] , \quad (C.13) \\
\mathbb{E} \left[ \tilde{Y}_i^{(p,p)} \right] & \leq \mathbb{E} \left[ b_{i1} \tilde{Y}_i^{(p-1,p)} \right] + \mathbb{E} \left[ b_{i2} \tilde{Y}_i^{(p-2,p)} \right] + \mathbb{E} \left[ b_{i3} \tilde{Y}_i^{(p-1,p-1)} \right] . \quad (C.14)
\end{align*}
\]

Here \( a_{ij}, b_{ij}, j = 1, 2, 3 \) are random variables satisfying that with some constants \( C_{p,K} \),

\[
|a_{i1}| \mathbb{E}[\tilde{\Xi}_i(z)] \leq N^{2\epsilon} \Psi^2, \quad |a_{i2}| \mathbb{E}[\tilde{\Xi}_i(z)] \leq N^{2\epsilon} \Psi^2, \quad |a_{i3}| I(\tilde{\Xi}_i(z)) \leq N^{2\epsilon} \Psi^2 , \quad \mathbb{E} |a_{ij}|^p \leq C_{p,K},
\]

for some event high probability event \( \tilde{\Xi}_i(z) \) such \( \mathbb{P}[\tilde{\Xi}_i(z)] > 1 - N^{-D_1} \), and the same bounds hold for \( \tilde{Y} \).

We then proceed to the proof of the Lemma C.2 assuming that both (C.13) and (C.14) hold. Similar to the discussion of (6.11), by applying Young’s inequality to (C.13), we find that for all \( \epsilon_1' > 0 \)

\[
\mathbb{E} \left[ a_{i1} \tilde{X}_i^{(p-1,p)} \right] \leq \frac{N^{2\epsilon \epsilon_1'}}{2p} \mathbb{E} \left[ |a_{i1}|^2 | \tilde{\Xi}_i(z) \right] + \frac{(2p-1)N^{-2\epsilon}}{2p} \mathbb{E} \left[ |\varphi(\Gamma_i)| P_i | \tilde{\Xi}_i(z) \right] .
\]

Utilizing the bound

\[
\mathbb{E} \left[ |a_{i1}|^2 \right] \leq \mathbb{E} \left[ |a_{i1}|^2 | I(\tilde{\Xi}_i(z)) \right] + \mathbb{E} \left[ |a_{i1}|^4 \right]^{1/2} \left( 1 - \mathbb{P}[\tilde{\Xi}_i(z)] \right)^{1/2} \leq N^{2\epsilon \epsilon_1'} \Psi^2 + C_{p,K} N^{-D_1/2},
\]

we get

\[
\mathbb{E} \left[ a_{i1} \tilde{X}_i^{(p-1,p)} \right] \leq C_{p,K} N^{2\epsilon \epsilon_1'} + C_{p,K} N^{-D_1/2+2\epsilon \epsilon_1'} + C_{p,K} N^{-2\epsilon_1'} \mathbb{E} \left[ |\varphi(\Gamma_i)| P_i | \tilde{\Xi}_i(z) \right] .
\]
Similarly, we can bound the terms regarding $\tilde{X}_i^{(p-2,p)}$ and $\tilde{X}_i^{(p-1,p)}$. Specifically, for $j = 2, 3$, we have that

$$\mathbb{E} \left[ \|a_{ij}\phi(\Gamma_i)P_i\|^2 \right] \leq C_{p,K} N^{D_1/2 + 2p' + 1} + C_{p,K} N^{-\frac{p+1}{2} - 2p'}. $$

Combining the above estimates, we arrive at

$$\mathbb{E} \left[ \|\phi(\Gamma_i)P_i\|^2 \right] \leq C_{p,K} N^{2p(1+\epsilon')} \Psi^{2p} + C_{p,K} N^{-D_1/2 + 2p' + 1}. \quad \text{(C.15)}$$

Therefore, by Markov inequality, we have

$$\mathbb{P} \left[ \|\phi(\Gamma_i)P_i\| > \Psi N^{\epsilon'/4} \right] \leq C_{p,K} N^{2p(1+\epsilon') - \Psi^{-2p} N^{-D_1/2 + 2p' + 1}} \quad \text{(C.16)}$$

We now require that $\epsilon, \epsilon' < \epsilon/16$, $p > (2D + 2)/\epsilon$ and $D_1 > 4p + 2D + 2$. Then (C.16) yields that

$$\mathbb{P} \left[ \|\phi(\Gamma_i)P_i\| > \Psi N^{\epsilon'/4} \right] \leq N^{-D_1}.$$  

Denote

$$\Xi_{11}(z) = \left\{ \|\phi(\Gamma_i)P_i\| \leq \Psi N^{\epsilon'/4} \right\}.$$

Since $\phi(\Gamma_i) = 1$ on $\Theta \left( z, \frac{N^{3\epsilon}}{(N\eta)^{1/3}}, \frac{N^{3\epsilon}}{N\eta} \right)$, we have $|P_i| \leq N^{\epsilon'/4} \Psi$ on the event $\Xi_{11}(z) \cap \Theta \left( z, \frac{N^{3\epsilon}}{(N\eta)^{1/3}}, \frac{N^{3\epsilon}}{N\eta} \right)$. Similarly, we can see that $|K_i| \leq N^{\epsilon'/4} \Psi$ on the event $\Xi_{12}(z) \cap \Theta \left( z, \frac{N^{3\epsilon}}{(N\eta)^{1/3}}, \frac{N^{3\epsilon}}{N\eta} \right)$, where $\Xi_{12}(z)$ is defined as

$$\Xi_{12}(z) = \left\{ \|\phi(\Gamma_i)K_i\| \leq \Psi N^{\epsilon'/4} \right\}.$$

This shows that we can further restrict ourselves on the high probability event $\Xi_{11}(z) \cap \Xi_{12}(z) \cap \Theta \left( z, \frac{N^{3\epsilon}}{(N\eta)^{1/3}}, \frac{N^{3\epsilon}}{N\eta} \right)$, where the following hold

$$P_i \leq N^{\epsilon'/4} \Psi, \quad K_i \leq N^{\epsilon'/4} \Psi.$$  

We can conclude the proof of Lemma 6.12 by setting $\Xi_1(z) : = \Xi_{11}(z) \cap \Xi_{12}(z)$ and following the proof of (6.25), i.e., the discussion between (6.26) and the paragraph before (6.30).

It remains to prove (C.13) and (C.14). We only discuss (C.13). In the proof of Lemma 6.4 we only require that $\Gamma_i$ to be $O_+(1)$. In this sense, the cutoff factor $\phi(\Gamma_i)$ can be used as an substitute for Assumption 6.2. Moreover, we can restrict our discussion on a high probability event using the property of Gaussian tail. Specifically, for any $\epsilon > 0$, let $\Xi_1'(z) \equiv \Xi_1'(z, \epsilon_1)$ be the event that all the large deviation estimates regarding the Gaussian vector $g_i$’s hold with precision with $N^\epsilon$. For example, the error bounds in (6.19), (6.13) and (6.15), which all are $O_+(N^{-1/2})$, can be replaced by $N^{-1/2+\epsilon_1}$ on $\Xi_1'(z)$. Using the Gaussian tail of $g_i$’s, we find that for any $D_1 > 0$, there exists an $N(D_1, \epsilon_1)$ such that for all $N \geq N(D_1, \epsilon_1)$, we have

$$\mathbb{P}(\Xi_1'(z)) \geq 1 - N^{-D_1}.$$  

Consequently, we can conclude that (C.13) has the same form as (6.38) except that cutoff factor $\phi(\Gamma_i)$. We can follow the proof of Lemma 6.4 and accommodate the additional terms in the integration by parts resulting from the derivatives of $\phi(\Gamma_i)$ into $X_i$. Note that the derivative of $\phi(\Gamma_i)$ appears in (6.50) and (6.71). For instance, in the analogue of (6.50), we will get an extra term

$$\frac{p}{N} \sum_{k}^{(i)} e_k^* B^{(i)} G e_i \|g_i\| \text{tr}(G A) \frac{\partial \phi(\Gamma_i)}{\partial g_{ik}} \Xi_i^{(p-1,p)}.$$
By replacing \( \varphi'(\Gamma_i) \) by \( CK^{-1} \) and chain rule, it suffices to replace the derivative of \( \varphi'(\Gamma_i) \) by derivatives of each of the followings:

\[
|G_{ii}|^2, \ |G_{ii}|^2, \ |T_i|^2, \ |\tilde{T}_i|^2, \ |\text{tr } G|^2, \ |\text{tr } (GA)|^2, \ |\text{tr } (BG)|^2.
\]

(C.17)

For an illustration, we provide the estimate only for \( |G_{ii}|^2 \). By definition, we only need to focus on the event \( \{\varphi'(\Gamma_i) \neq 0\} \cap \Xi'_i(z) \). Using

\[
\frac{\partial |G_{ii}|^2}{\partial g_{ik}} = \frac{\partial G_{ii}}{\partial g_{ik}} + \frac{\partial G_{ii}}{\partial g_{ik}} g_{ik},
\]

the corresponding term becomes

\[
\frac{G_{ii}}{N} \left( \sum_{k}^{(i)} e_k^* \tilde{B}^{(i)} G e_i \frac{\partial G_{ii}}{\partial g_{ik}} (\tilde{\Xi}'_{(p-1,p)}) + \frac{G_{ii}}{N} \text{tr}(GA) \right) \sum_{k}^{(i)} e_k^* \tilde{B}^{(i)} G e_i \frac{\partial G_{ii}}{\partial g_{ik}} \tilde{\Xi}'_{(p-1,p)}.
\]

Note that the above terms can be controlled using Lemma B.3 and in fact,

\[
\frac{p}{N} \sum_{k}^{(i)} e_k^* \tilde{B}^{(i)} G e_i \frac{\partial G_{ii}}{\partial g_{ik}} \tilde{\Xi}'_{(p-1,p)} = O \left( \frac{N^{r_1}}{\sqrt{N\eta}} \right).
\]

The rest terms in (C.17) can be handled in the same way and we omit the details here. In summary, the additional terms regarding the derivatives of \( \varphi(\Gamma_i) \) can be absorbed into the first term on the RHS of (C.13). This completes our proof.

Finally, we prove Lemma C.3. We need to handle extra technical complexity compared to the proof of Lemma C.2. For some sufficiently small constant \( c > 0 \), denote

\[
\Gamma = \left( \frac{N^{5c}}{\sqrt{N\eta}} \right) ^{-1} \frac{1}{N} \sum_i |T_i|^2 + N^{-1/2} + \left( \frac{N^{5c}}{N\eta^{1/2}} \right) ^{-2} |\Gamma|^2 \left( z \text{ tr}(GA) - \Omega_B(zm_G(z) + 1) + |z \text{ tr}(BG) - \Omega_A(zm_G(z) + 1)|^2 + |zG + 1 - (zm_G(z) + 1)|^2 \right)
\]

\[
+ \left( c \text{ Im } m_G(z) + \tilde{A} \right) ^2,
\]

where used the short-hand notation \( m_G = m_{\mu_B} \). Recall (6.12) and (7.44). Denote

\[
\tilde{\Xi}'_{(p,q)} = \left( \frac{1}{N} \sum_i d_i \varphi(\Gamma_i) \varphi(\Gamma) Q_i \right)^p \left( \frac{1}{N} \sum_i d_i \varphi(\Gamma_i) \varphi(\Gamma) Q_i \right)^q, \quad d_i = \tilde{d}_{1i} \text{ or } \tilde{d}_{12}.
\]

**Proof of Lemma C.3** Similar to the construction of the event \( \Xi'_i(z) \) in the proof of Lemma C.2 for any \( \epsilon > 0 \), we denote \( \Xi'_i(z) \) be the event such that all the large deviation estimates regarding the Gaussian random vectors \( g_i \)’s in the proof of Lemma C.3 hold with precision \( N' \). Again, by the properties of Gaussian tails, we find that for any \( D > 0 \), there exists an \( N'(D, \epsilon) \) such that when \( N \geq N'(D, \epsilon) \), the following holds

\[
\mathbb{P}(\Xi'_i(z)) \geq 1 - N^{-D}.
\]

Analogously to the proof of Lemma C.2 (c.f. (C.13) and (C.14)), it suffices to prove the following recursive moment estimate when \( N \) is sufficiently large

\[
\mathbb{E} \left[ \tilde{\Xi}'_{(p,q)} \right] \leq \mathbb{E} \left[ \epsilon_1 \tilde{\Xi}'_{(p-1,p)} \right] + \mathbb{E} \left[ \epsilon_2 \tilde{\Xi}'_{(p-2,p)} \right] + \mathbb{E} \left[ \epsilon_3 \tilde{\Xi}'_{(p-1,p-1)} \right],
\]

(C.19)

where \( \epsilon_j \) for \( j = 1, 2, 3 \) are random variables satisfying that

\[
|\epsilon_1| \mathbb{E}(\Xi'_i(z)) \leq N^\alpha \hat{\Pi}, \quad |\epsilon_2| \mathbb{E}(\Xi'_i(z)) \leq N^{2\alpha} \hat{\Pi}^2, \quad |\epsilon_3| \mathbb{E}(\Xi'_i(z)) \leq N^{2\alpha} \hat{\Pi}^2, \quad \mathbb{E} |\epsilon_j|^p \leq C_{p,k}.
\]

83
We mention that, unlike that \([\text{C.13}]\) is a direct cutoff version of \([\text{C.3}]\), \([\text{C.19}]\) is a cutoff to \([\text{C.3}]\) with a weaker bounds for the coefficients \(c_i\)'s. The weakness of the bounds is due to the weak a priori inputs in the cutoffs \(\varphi(\Gamma_1)\) and \(\varphi(\Gamma)\), and the terms involving the derivatives of these cutoffs. We then discuss both of these two aspects and only list the necessary modifications.

First, we discuss the technical inputs in the cutoffs. Before explaining the detailed modifications, we introduce some useful estimates. We observe that

\[
\Omega_B(z)(zm_{\mu_A}zg_{\mu_B}(z) + 1) = \Omega_B(z)(\Omega_B m_{\mu_A}(\Omega_B(z) + 1) = \int \frac{x^2}{x - \Omega_B(z)} d\mu_A(x) - 1,
\]

where in the second step we used the assumption that \(\text{tr} A = 1\) and the identity

\[
z(zm_{\mu}(z) + 1) = \int \frac{x^2}{x - z} d\mu(x) - \int x d\mu(x).
\]

In what follows, we will write \(\Omega_B \equiv \Omega_B(z)\) for simplicity. Based on \([\text{C.20}]\), we find that

\[
\text{Im} (\Omega_B(zm_{\mu_A}zg_{\mu_B}(z) + 1)) = \text{Im} \Omega_B \int \frac{x^2}{|x - \Omega_B|^2} d\mu_A(x) \geq a_N \text{Im} \Omega_B \int \frac{x}{|x - \Omega_B|^2} d\mu_A(x)
\]

where we used \(\text{inf sup} \mu_A \equiv \mu_B \geq a_N b_N\) in the last inequality. Similarly, by reversing the inequities in \([\text{C.21}]\), we have \(\text{Im} \Omega_B(zm_{\mu_A}zg_{\mu_B}(z) + 1)) \leq \frac{a_1 b_1}{b_N} \text{Im} m_{\mu_A}zg_{\mu_B}(z)\). Hence, we have that for some constants \(c_K, c_K' > 0\), when \(c\) is sufficiently small

\[
\varphi(\Gamma)|\text{tr} G A| \geq \varphi(\Gamma) \left(\Omega_B(zm_{\mu_A}zg_{\mu_B}(z) + 1)) - \sqrt{2K}(c \text{Im} m_{\mu_A}zg_{\mu_B}(z) + \hat{\Lambda})\right)
\]

\[
\geq \frac{\sqrt{2}}{2} \varphi(\Gamma) \left[|\text{Re} \Omega_B(zm_{\mu_A}zg_{\mu_B}(z) + 1))| + \frac{a_N b_N}{a_1^2 b_1} \left(1 - \frac{c\sqrt{2K}}{a_N^2 b_N} \right) \text{Im} m_{\mu_A}zg_{\mu_B}(z) - c\sqrt{2K} N^{-\gamma/3}\right]
\]

\[
\geq \frac{c_K}{2} \varphi(\Gamma) \left[|\text{Re} \Omega_B(zm_{\mu_A}zg_{\mu_B}(z) + 1))| + \frac{a_N b_N}{a_1^2 b_1} \left(1 - \frac{c\sqrt{2K}}{a_N^2 b_N} \right) \text{Im} (\Omega_B(zm_{\mu_A}zg_{\mu_B}(z) + 1))\right]
\]

\[
\geq \frac{c_K}{4} \varphi(\Gamma) |\Omega_B(zm_{\mu_A}zg_{\mu_B}(z) + 1))|
\]

\[
\geq c'_K \varphi(\Gamma),
\]

where in the first step we used the definitions of \(\varphi(\cdot)\) in \([\text{C.11}]\) and \(\Gamma\) in \([\text{C.18}]\), in the second step we used \(x^2 + y^2 \geq (x + y)^2/2\), \([\text{C.21}]\), the definition \([\text{C.3}]\) and the fact \(\eta \geq N^{-1+\gamma}\), in the third step we used the assumption \(\epsilon < \gamma/12\), in the fourth step we used \(x + y \geq \sqrt{(x^2 + y^2)/2}\) when \(x, y \geq 0\) and the assumption \(c\) is sufficiently small and in the last step we used a relation similar to \([\text{C.10}]\) and (i) of Proposition 3.1. Similarly, we can show that

\[
\varphi(\Gamma)|\text{tr} (zG + 1))| \geq c'_K \varphi(\Gamma).
\]

Analogously, for some constant \(C_K > 0\), we have that

\[
\varphi(\Gamma)|\text{tr} G \sim \varphi(\Gamma)|\text{tr} (zG + 1) \sim \varphi(\Gamma)|\text{tr} (GA)
\]

\[
\leq \varphi(\Gamma)C_K(\text{Im} (zm_{\mu_A}zg_{\mu_B}(z) + 1)) + \hat{\Lambda}) \sim \varphi(\Gamma)C_K(\text{Im} m_{\mu_A}zg_{\mu_B}(z) + \hat{\Lambda}).
\]

This implies that

\[
\varphi(\Gamma) \frac{N^\epsilon}{N^2} \sum_i |G_{ii}| \leq \varphi(\Gamma) \frac{N^\epsilon}{N} \frac{\text{Im} \text{tr} G}{\eta} \leq \varphi(\Gamma)C_K \frac{N^\epsilon}{\sqrt{N}} \sqrt{\frac{\text{Im} m_{\mu_A}zg_{\mu_B}(z) + \hat{\Lambda}}{N\eta}} \leq \frac{N^\epsilon}{\sqrt{N}} \hat{\Gamma}.
\]

(2.24)
Moreover, by the definition of \( \Gamma \), we have that
\[
\varphi(\Gamma) \frac{1}{N} \sum_i |T_i| \leq \varphi(\Gamma) \frac{1}{N} \sum_i (|T_i|^2 + N^{-1})^{1/2} \leq 2K \varphi(\Gamma) \frac{N^{5\epsilon}}{\sqrt{N\eta}}.
\] (C.25)

This yields that
\[
\varphi(\Gamma) \frac{N^{5\epsilon}}{N^{3/2}} \sum_i |T_i| \leq \frac{N^{7\epsilon}}{\sqrt{N^{2\eta}}} \leq N^{\epsilon} \hat{\Pi}.
\] (C.26)

With above preparation, we explain the modifications regarding the technical inputs. The first quantity is the \( \tau_{i1} \) defined in (7.9). The counterpart of \( \tau_{i1} \) is the following
\[
\tilde{\tau}_{i1} := a_i \frac{\text{tr}(G\tilde{D})}{\text{tr}(GA)} - \tilde{d}_i, \quad \tilde{D} = \text{diag}(\tilde{d}_1, \ldots, \tilde{d}_N), \quad \tilde{d}_i := d_i \varphi(\Gamma_i) \varphi(\Gamma).
\] (C.27)

By (C.22), (C.23) and the fact \( \varphi(\Gamma_i)|G_{ii}| \leq 2K \), we conclude \( \tilde{\tau}_{i1} \leq C_K \) for some constant \( C_K > 0 \). The second modification is to discuss the bound of the analogue of the term \( O_\omega(\Psi\hat{\Pi}) \) in (6.4). This error term was obtained when we bound \( \frac{1}{N} \sum_i T_i \chi_{\tau_{i1}} \) in (7.10). For its counterpart in the current proof, using the definition of \( \Gamma \) in (C.18) and the bound in (C.25), we conclude that for some constant \( C > 0 \)
\[
\varphi(\Gamma) \frac{1}{N} \sum_i T_i \chi_{\tau_{i1}} \leq C \varphi(\Gamma) C_K \frac{N^{5\epsilon}}{\sqrt{N\eta}} \frac{N^{5\epsilon}}{(N\eta)^{1/3}} \leq C \varphi(\Gamma) C_K \frac{N^{10\epsilon}}{(N\eta)^{5/6}} \leq \sqrt{\frac{\Lambda}{N\eta}} \leq \hat{\Pi},
\]
where we recall the definitions in (C.4). This implies that \( \frac{1}{N} \sum_i T_i \chi_{\tau_{i1}} \) can be absorbed into the bound for \( \epsilon_1 \) in (C.19). The third and fourth modifications are devoted to the terms involving \( e_{i1} \) in (6.42) and \( e_{i2} \) in (6.67). By a discussion similar to (7.18), we have that
\[
\varphi(\Gamma) e_{i1} = (-h_{ii} + \hat{h}_i^* (\hat{B}^{(i)} - I) \hat{h}_i) G_{ii} + O_\omega(1/N) G_{ii} + O_\omega(1/\sqrt{N}) T_i.
\]
Without loss of generality, we assume that the stochastic dominance holds on some event \( \Xi_2 \) with parameters \( \epsilon \) and \( D \). Then taking average over \( i \) and focusing on the above event \( \Xi_2 \), we have
\[
\varphi(\Gamma) \frac{1}{N} \sum_i e_{i1} \tilde{\tau}_{i1} \text{tr}(GA) = \varphi(\Gamma) \frac{1}{N} \sum_i (-h_{ii} + \hat{h}_i^* (\hat{B}^{(i)} - I) \hat{h}_i) G_{ii} \tilde{\tau}_{i1} \text{tr}(GA) + O(N^{\epsilon} \hat{\Pi}),
\] (C.28)
where we used (C.24) and (C.26). Similarly, we have that
\[
\varphi(\Gamma) \frac{1}{N} \sum_i \frac{e_{i2} \tilde{\tau}_{i1}}{\|g_i\|} = \varphi(\Gamma) \frac{1}{N} \sum_i (g_i^* g_i - 1) \tilde{\tau}_{i1} \text{tr}(A(\hat{B} - I)G) - h_{ii} \text{tr}(GA)) G_{ii} + O(N^{\epsilon} \hat{\Pi}).
\] (C.29)

Similar to the discussion in (7.21) and (7.22), the second term of (C.24) cancels with the first term of (C.28), and the rest will be dealt with the technique of integration by parts.

Second, we estimate are additional terms arising from the derivatives of \( \varphi(\Gamma) \) and \( \varphi(\Gamma_i) \). Since the terms involving derivatives of \( \varphi(\Gamma_i) \) can be handled in the same way as in the proof of Lemma C.2, we only focus on the discussion of \( \varphi(\Gamma) \). In light of (C.27), the derivatives containing \( \varphi(\Gamma) \) appear in the analogues of (7.14), (7.17), and the new terms after the integration by parts using the following two terms from (C.28) and (C.29)
\[
\frac{1}{N} \sum_i \mathbb{E} \left[ \hat{h}_i^* (\hat{B}^{(i)} - I) \hat{h}_i G_{ii} \tilde{\tau}_{i1} \text{tr}(GA) \tilde{X}^{(p-1,p)} \right], \quad \frac{1}{N} \sum_i \mathbb{E} \left[ (\hat{g}_i^* \hat{g}_i - 1) \tilde{\tau}_{i1} \text{tr}(A(\hat{B} - I)G) G_{ii} \tilde{X}^{(p-1,p)} \right].
\]

85
Due to similarity, we will only focus on discussing the counterpart of the analogue of the step (C.33) and the other terms can be handled similarly. More specifically, we study

$$\frac{1}{N^2} \sum_i \sum_k e_i^* \tilde{B}^{(i)}(G) e_k \frac{\partial \tau_{11}}{\partial g_{ik}} \text{tr}(GA).$$  \tag{C.30}$$

Recall (C.27). One new term in $\partial \tau_{11}/\partial g_{ik}$ is

$$-d_i \psi(\Gamma_i) \psi'(\Gamma) \frac{\partial \Gamma}{\partial g_{ik}}. \tag{C.31}$$

In what follows, we study the contribution of the term (C.31) to (C.30). The other terms can be handled similarly. In view of (C.18) and (C.31), it remains to analyze the derivative of each term in $\Gamma$. In the sequel, we only focus on the term involving $([T_j]^2 + N^{-1})$ and $\Gamma$. The other terms can be investigated in the same way (actually easier) since they are tracial.

Recall (6.13). For the term involving $([T_j]^2 + N^{-1})$, using

$$\frac{\partial (e^{i\theta} T_j)}{\partial g_{ik}} = \frac{\partial (e_i^* U G e_j)}{\partial g_{ik}} = e_i^* U ( \frac{\partial (G^* R_i)}{\partial g_{ik}} ) e_j,$$

we find that

$$\frac{\partial ([T_j]^2 + N^{-1})}{\partial g_{ik}} = \frac{1}{([T_j]^2 + N^{-1})^{1/2}} \left( e_i^* G^* U e_j e_i^* U ( \frac{\partial (R_i G)}{\partial g_{ik}} ) e_j + e_i^* \frac{\partial (G^* R_i)}{\partial g_{ik}} (U^{(i)})^* e_j e_j^* U G e_j \right). \tag{C.32}$$

Recall (6.31) and (6.33). It is easy to see that

$$\frac{\partial R_i G}{g_{ik}} = \frac{2}{\||g_i\||} (e_i^* (e_i + h_i)^* G + R_i G A e_k (e_i + h_i)^* \tilde{B}^{(i)} R_i G + R_i G A \tilde{B}^{(i)} e_k (e_i + h_i)^* G)$$

$$+ \Delta R(i, k) G + R_i \Delta G(i, k). \tag{C.33}$$

Based on the above expression, the definition of $T_j$ in (6.13) and (C.32), the main contribution of the first term in (C.33) to (C.31) and (C.30) involving $([T_j]^2 + N^{-1})$ reads

$$\left( \frac{N^5}{\sqrt{\nu N \eta}} \right)^{-1} \frac{1}{N^3} \sum_{i,j} e_i^* \left( \frac{e_i^* T_j^*}{(\|i\|_2 + N^{-1})^{1/2}} \frac{\ell_i^2}{\|g_i\|^2} \right) e_i^* \tilde{B}^{(i)}(G) e_i \text{tr}(GA) e_j^* U^{(i)} e_k (e_i + h_i)^* G e_j$$

$$= \left( \frac{N^5}{\sqrt{\nu N \eta}} \right)^{-1} \frac{1}{N^3} \sum_i \ell_i^2 \|g_i\|^2 \text{tr}(GA) (e_i + h_i)^* G \mathcal{D}_1 U^{(i)} I_i \tilde{B}^{(i)} G e_i, \tag{C.34}$$

where $I_i := I - e_i e_i^*$ and

$$\mathcal{D}_1 := \text{diag} \left( \frac{e_i^* T_j^*}{(\|i\|_2 + N^{-1})^{1/2}} \right)_{j=1,\ldots,N}.$$  

Then we can follow the proof of Lemma B.3 to get that for some constant $C > 0$

$$\left| (e_i + h_i)^* G \mathcal{D}_1 U^{(i)} I_i \tilde{B}^{(i)} G e_i \right| \leq C_K \left( \|G^* e_i\|^2 + \|G^* h_i\|^2 + \|G e_i\|^2 \right) \leq C \frac{\text{Im} G_{ii} + \text{Im} \tilde{G}_{ii}}{\eta},$$

where we used that

$$\|G^* h_i\|^2 = h_i^* G G^* h_i = e_i^* U^* G G^* U e_i = e_i^* G G^* e_i = \frac{1}{b_i} e_i^* \tilde{G} B \tilde{G}^* e_i.$$  

86
To sum up, we have

$$|\text{(C.34)}| \leq C \frac{1}{N} \left( \frac{N^4 \epsilon}{\sqrt{N} \eta} \right)^{-1} \tilde{\Pi}^2 \leq \tilde{\Pi}^2.$$  

The same reasoning also applies to the remaining terms of \text{(C.33)}, and we thus conclude the estimate for the derivative involving \(|T_j|^2 + N^{-1})^{1/2}.

We finally discuss the terms involving the derivatives of \(\Upsilon\), which is defined in \text{(6.16)}. By the definition of \(\Upsilon\), we have that

$$\frac{\partial \Upsilon}{\partial g_{ik}} = z \left( \text{tr} \left( \frac{\partial G}{\partial g_{ik}} \right) A \text{tr}(ABG) + \text{tr} \left( \frac{\partial G}{\partial g_{ik}} \right) \text{tr}(ABG) - \text{tr}(G) \text{tr}(AG) \right),$$  

(C.35)

and that

$$\frac{\partial |\Upsilon|^2}{\partial g_{ik}} = \frac{\partial \Upsilon}{\partial g_{ik}} \Upsilon + \Upsilon \frac{\partial \Upsilon}{\partial g_{ik}}.$$  

(C.36)

Due to similarity, for \text{(C.35)}, we only discuss the contribution of \(\Upsilon \frac{\partial \Upsilon}{\partial g_{ik}}\) to \text{(C.31)}. Specifically, we study

$$\left( \frac{N^{5e}}{(N \eta)^{1/3}} \right)^{-2} \frac{T}{N^2} \sum_i \sum_k \frac{1}{\|g_i\|} e^{i_k \tilde{B} (i)} \frac{\partial \Upsilon}{\partial g_{ik}} \text{tr}(GA),$$

which is reduced to bound all the terms in \text{(C.35)}. For instance, for the first term in \text{(C.35)}, we have that for some constants \(C, C_K > 0\)

$$\left| \left( \frac{N^{5e}}{(N \eta)^{1/3}} \right)^{-2} \frac{T}{N^2} \sum_i \sum_k \frac{1}{\|g_i\|} e^{i_k \tilde{B} (i)} \frac{\partial \Upsilon}{\partial g_{ik}} \text{tr}(GA) \right| \leq \varphi(\Gamma) C_K \left( \frac{N^{5e}}{(N \eta)^{1/3}} \right)^{-1} \frac{1}{N^2} \sum_i \sum_k e^{i_k \tilde{B} (i)} \frac{\partial \Upsilon}{\partial g_{ik}} \text{tr}(GA) \leq C_K \varphi(\Gamma) \left( \frac{N^{5e}}{(N \eta)^{1/3}} \right)^{-1} \frac{1}{N} \Pi^2 \eta^2 \leq \Pi^2,$$

where in the first step we used the definition of \(\Gamma\) and \(\varphi(.)\). Similar argument applies to the other terms in \text{(C.35)}.

Except for the aforementioned changes, the rest of the proof of \text{(C.19)} is the same as that for \text{(7.3)}. This concludes the proof of Lemma \text{(C.8)}.

\section{Additional technical proofs}

\subsection{Proofs of Proposition \text{3.9}, Lemmas \text{3.12}, \text{3.13} and \text{3.14}}

\textbf{Proof of Proposition \text{3.9}} First of all, we choose \(\tau\) and \(\eta\) to satisfy the conclusions of Lemmas \text{3.6} and \text{3.8}. We start with the proof of (i). The last two equivalences follow from an elementary computation by taking the imaginary parts of \text{(A.13)} and using the fact that \(\Omega_\alpha(E_+), \Omega_\beta(E_+) \in \mathbb{R}^+\). By \text{(2.10)} and (ii) of Lemma \text{3.2} we find that

$$\text{Im}(zm_{\mu_\alpha \otimes \mu_\beta}(z)) = \eta \int \frac{x}{|x - z|^2} d(m_{\mu_\alpha \otimes \mu_\beta})(x) \sim \eta \int \frac{1}{|x - z|^2} d(m_{\mu_\alpha \otimes \mu_\beta})(x) = \text{Im} m_{\mu_\alpha \otimes \mu_\beta}(z).$$  

(D.1)
Moreover, by the definition of $M$-transform in (3.4), we find that
\[
\text{Im} M_{\mu_\alpha, \mu_\beta} = \frac{\text{Im}(zm_{\mu_\alpha, \mu_\beta}(z))}{|1 + zm_{\mu_\alpha, \mu_\beta}(z)|^2} \sim \text{Im}(zm_{\mu_\alpha, \mu_\beta}(z)),
\]
where we again use (2.10) and (ii) of Lemma 3.2. Together with (D.1), we conclude that
\[
\text{Im} M_{\mu_\alpha, \mu_\beta} \sim \text{Im} m_{\mu_\alpha, \mu_\beta}(z).
\]
By (2.10), we find that
\[
\text{Im} zm_{\mu_\alpha, \mu_\beta}(z) = \text{Im} \Omega_\beta(z)m_{\mu_\alpha}(\Omega_\beta(z)) = \text{Im} \Omega_\beta(z) \int \frac{x}{|x - \Omega_\beta(z)|^2} \text{d}\mu_\alpha(x).
\]
Together with (ii) of Lemma 3.2, we obtain that
\[
\text{Im} zm_{\mu_\alpha, \mu_\beta}(z) \sim \text{Im} \Omega_\beta(z).
\]

Similarly, we find that
\[
\text{Im}(zm_{\mu_\alpha, \mu_\beta}(z)) = \eta \int \frac{x}{|x - z|^2} \text{d}(\mu_\alpha \otimes \mu_\beta)(x) \sim \eta \int \frac{1}{|x - z|^2} \text{d}(\mu_\alpha \otimes \mu_\beta)(x) = \text{Im} m_{\mu_\alpha, \mu_\beta}(z).
\]
This concludes the proof of (i).

Then we prove (ii) and start with the first term. First, when $z \in D_\tau(0, \eta_\beta)$, by differentiating $\tilde{z}_\tau$ defined in (3.11) with respect to $\Omega$, we find that
\[
-M'_{\mu_\alpha}(\Omega_\beta(z))\tilde{z}_\tau'(\Omega_\alpha(z)) = -M'_{\mu_\alpha}(\Omega_\beta(z)) \left( \frac{\Omega_\beta(z)}{M_{\mu_\alpha}(\Omega_\beta(z))} + \frac{\Omega_\alpha(z)}{M_{\mu_\alpha}(\Omega_\alpha(z))} M'_{\mu_\alpha}(\Omega_\alpha(z)) - \Omega_\alpha(z)\Omega_\beta(z) M_{\mu_\alpha}(\Omega_\alpha(z))^2 \right),
\]
where in the first equality we use (2.3) and the inverse function theorem. Note that $M_{\mu_\alpha}^{-1}(\cdot)$ is analytic around $M_{\mu_\beta}(\Omega_\alpha(z))$ since $M_{\mu_\alpha}(\Omega_\beta(z)) \sim 1$. By Remark 3.4 and (D.2), we readily find that
\[
-M'_{\mu_\alpha}(\Omega_\beta(z))\tilde{z}_\tau'(\Omega_\alpha(z)) = S_{\alpha\beta}(z),
\]
By (3.10), we find that
\[
M'_{\mu_\alpha}(\Omega_\beta(z)) = 1 + \int \frac{x}{(x - z)^2} \text{d}\mu_\alpha(x).
\]
Since $\text{supp} \, \mu_\alpha = \text{supp} \, \mu_\alpha$, by (ii) of Lemma 3.2, we find that $M'_{\mu_\alpha}(\Omega_\beta(z)) \sim 1$. Therefore, the first term of (ii) follows from (D.3), (3.18) and (3.13) when $z \in D_\tau$. Second, when $z \in D_\tau(0, \eta_\beta) \setminus D^* = D_\tau(\eta_0, \eta_\beta)$, following the a discussion similar to the proof of (27) Lemma 6.6, we get
\[
|S_{\alpha\beta}(z)| \geq 1 - |S_{\alpha\beta}(z) + 1| \geq 1 - |z|^2 \left( \int \frac{1}{|x - \Omega_\alpha(z)|^2} \text{d}\mu_\alpha(x) \right) \left( \int \frac{1 - |z|^2 \text{Im} \Omega_\alpha(z)/z}{\text{Im} \Omega_\beta(z)/z} \text{Im} \Omega_\beta(z)/z \right) \geq 1 - |z|^2 \frac{\text{Im} \Omega_\alpha(z)/z}{\text{Im} \Omega_\beta(z)/z} \geq c > 0,
\]
where we use the fact that $\eta_0 = O(1)$. This completes the proof for $S_{\alpha\beta}(z)$. We then prove the second and third terms of (ii). Due to similarity, we only discuss how to bound $T_\alpha(z)$. Recall the definition of $T_\alpha(z)$ in
Since \( z \in \mathcal{D}_r(0, \eta_U) \) is bounded, it suffices to control \( L'_\mu_{\beta}(\Omega_\alpha(z)) \) and \( L''_{\mu_{\beta}}(\Omega_\alpha(z)) \). By (3.10), we find that

\[
L'_{\mu_{\beta}}(\Omega_\alpha(z)) = \int \frac{1}{(x - \Omega_\alpha(z))^2} d\hat{\mu}_{\beta}(x), \quad L''_{\mu_{\beta}}(\Omega_\alpha(z)) = \int \frac{1}{(x - \Omega_\alpha(z))^3} d\hat{\mu}_{\beta}(x).
\]

Hence, we can conclude the proof using \( \text{supp} \hat{\mu}_\alpha = \text{supp} \mu_\alpha \) and (ii) of Lemma 3.2.

Finally, we prove (iii) and only provide the details for \( \mathcal{T}_\alpha \). Since we have provided a upper bound in (ii), it suffices to find a lower bound. By (iii) of Lemma 3.2 that \( \Omega_\alpha(E_+) > E_+^3, \Omega_\beta(E_+) > E_+^\mu \) and Lemma 3.3 we have

\[
L'_{\mu_{\beta}}(\Omega_\alpha(E_+)) = \int \frac{1}{(x - \Omega_\alpha(E_+))^2} d\hat{\mu}_{\beta}(x) > 0, \quad L''_{\mu_{\beta}}(\Omega_\alpha(E_+)) = \int \frac{1}{(x - \Omega_\alpha(E_+))^3} d\hat{\mu}_{\beta}(x) < 0.
\]

Consequently, \( \mathcal{T}_\alpha(E_+^\beta) < 0 \) and \( \mathcal{T}_T(E_+) < 0 \). Since \( \Omega_\alpha, \Omega_\beta \) are continuous around \( E_+ \) and \( L_{\mu_{\beta}} \) are analytic around \( \Omega_\beta(E_+) \) and \( \Omega_\alpha(E_+) \) respectively, we find that the bounds hold in a small neighborhood of \( E_+ \). This concludes our proof.

**Proof of Lemma 3.12** We present only the proof for \( \mu_A \) and the proof for \( \mu_B \) is exactly the same. Denote the cumulative distribution functions of \( \mu_\alpha \) and \( \mu_A \) by \( F_\alpha \) and \( F_A \). We take \( N_0 \) such that \( \mathbf{d} \leq \delta/2 \) and \( [a_N, a_1] \in [E_+^\alpha - \delta/2, E_+^\alpha + \delta/2] \) for all \( N \geq N_0 \). Then By the definition of \( \mathcal{L}(\cdot, \cdot) \), we have

\[
|F_\alpha(x) - F_A(x)| \leq 1_{[E_+^\alpha - \delta/2, E_+^\alpha + \delta/2]}(x) \left( F_\alpha(x + \mathbf{d}) - F_\alpha(x - \mathbf{d}) + 2\mathbf{d} \right).
\]

Thus integrating by parts we get

\[
\left| \int_{\mathbb{R}_+} f(x)d(\mu_\alpha - \mu_A)(x) \right| = \left| \int_{E_+^\alpha - \delta/2}^{E_+^\alpha + \delta/2} f'(x)(F_\alpha(x) - F_A(x))dx \right|
\leq 2\mathbf{d}(E_+^\alpha - E_+^\alpha + \delta) \sup_{x \in [E_+^\alpha - \delta/2, E_+^\alpha + \delta]} |f'(x)| + \left| \int_{E_+^\alpha - \delta/2}^{E_+^\alpha + \delta/2} f'(x)(F_\alpha(x + \mathbf{d}) - F_\alpha(x - \mathbf{d}))dx \right|
\leq 2\mathbf{d}(E_+^\alpha - E_+^\alpha + \delta) \sup_{x \in [E_+^\alpha - \delta/2, E_+^\alpha + \delta]} |f'(x)| + \left| \int_{E_+^\alpha} f'(x)(|F_\alpha(x + \mathbf{d})| - |F_\alpha(x + \mathbf{d})|) F_\alpha(x)dx \right|
\leq 2\mathbf{d}(E_+^\alpha - E_+^\alpha + \delta) \| f' \|_{\text{Lip}, \delta}.
\]

This concludes the proof.

**Proof of Lemma 3.13** The proof follows from Lemma 3.6. More specifically, by Lemma 3.6 we find that there exist \( \eta_1 > 0 \) and \( \theta \in (0, \pi/2) \) such that for all \( z \in \mathcal{E}(\eta_1, \theta) \)

\[
|\Omega_A(z) - \Omega_\alpha(z)| \leq 2 \|r(z)\|, \quad |\Omega_B(z) - \Omega_\beta(z)| \leq 2 \|r(z)\|.
\]

Since

\[
\{ z \in \mathbb{C}_+: \text{Re} z \in [E_+ - \tau, \tau^{-1}], \text{Im} z > \eta_1 \} \subset \mathcal{E}(\eta_1, \theta),
\]

for large constant \( \eta > \eta_1 \), we conclude the proof.

It remains to verify the conditions of Lemma 3.6. By Lemma 3.11 and (v) of Assumption 2.4 we find that there exists a large constant \( \tilde{\eta}_1 > 0 \), such that for \( z = E + i\tilde{\eta}_1 \)

\[
|m_{\tilde{\mu}_A}(z)| \leq \frac{1}{2}, \quad |m_{\tilde{\mu}_B}(z)| \leq \frac{1}{2}.
\]  \hspace{1cm} (D.4)
Recall (B.19) By Lemma 3.12, we find that \( \Omega_A, \Omega_B : \mathcal{E}(\eta_1, 0) \to \mathbb{C}_+ \) are analytic functions. We now apply Lemma B.3 with the choice \((\mu_1, \mu_2) = (\mu_{\alpha}, \mu_{\beta})\), \( (\eta_1, \eta_2) = (\Omega_A, \Omega_B), \Phi = \Phi_{\alpha\beta} \) and \( e = 1/2 \). The first two conditions of (B.20) hold by (3.24) and (D.4). To verify the third condition of (B.20), we find that

\[
|r_A(z)| = |\Omega_B(z)|^{-1} \left| \int \frac{x}{x - \Omega_B(z)} d\mu_A(x) \right| \left| \int \frac{x}{x - \Omega_B(z)} d\mu_A(x) \right|^{-1} \left| \int \frac{xd(\mu_A - \mu_\alpha)(x)}{x - \Omega_B(z)} \right|. \tag{D.5}
\]

By (D.20), we have \( \text{Im} \Omega_B(z) \geq \eta_1 \) for \( z \in \mathcal{E}(\eta_1, 0) \). Together with Lemma 3.12 for some constant \( C \)

\[
\left| \int \frac{x}{x - \Omega_B(z)} d(\mu_A - \mu_\alpha)(x) \right| \leq C d. \tag{D.6}
\]

Further, with the same reasoning that \( \text{Im} \Omega_B(z) \geq \eta_1 \), it is easy to see that for some constant \( c_0 > 0 \),

\[
\left| \int \frac{x}{x - \Omega_B(z)} d\mu_A(x) \right| \geq c_0, \quad \left| \int \frac{x}{x - \Omega_B(z)} d\mu_A(x) \right| \geq c_0. \tag{D.7}
\]

By (D.6) and (D.7), we find that for some constant \( C > 0 \), we have \( |r_A(z)| \leq C d \). Similarly, we have that \( |r_B(z)| \leq C d \). This implies that

\[
\|r(z)\| \leq cd. \tag{D.8}
\]

Invoking (iv) of Assumption 2.4 we conclude that the third condition of (B.20) holds when \( N \) is large enough. \( \square \)

**Proof of Lemma 3.14** By (2.8), we have that

\[
L_{\mu_\alpha}(\Omega_B(z_0)) = L_{\mu_\alpha}(\Omega_\beta(z_0)) + \frac{\Delta \Omega_1(z_0)}{z_0} + r_A(z_0), \tag{D.9}
\]

where we denote

\[
\Delta \Omega(z_0) = (\Delta \Omega_1(z_0), \Delta \Omega_2(z_0)) := (\Omega_A(z_0) - \Omega_\alpha(z_0), \Omega_B(z_0) - \Omega_\beta(z_0)).
\]

Expanding \( L_{\mu_\alpha}(\Omega_B(z_0)) \) around \( \Omega_\beta(z_0) \), we have

\[
L_{\mu_\alpha}(\Omega_B(z_0)) = L_{\mu_\alpha}(\Omega_\beta(z_0)) + L'_{\mu_\alpha}(\Omega_\beta(z_0)) \Delta \Omega_2(z_0) + L''_{\mu_\alpha}(\xi_0) \Delta \Omega_2(z_0)^2, \tag{D.10}
\]

where \( \xi_0 \) is some value between \( \Omega_B(z_0) \) and \( \Omega_\beta(z_0) \). Note that

\[
\inf_{x \in \text{supp} \mu_\alpha} |x - \xi_0| \geq \inf_{x \in \text{supp} \mu_\alpha} |x - \Omega_\beta(z_0)| - |\Omega_\beta(z_0) - \xi_0| \geq \kappa_0 - \delta - q \geq \kappa_0/3,
\]

where \( \delta \) is defined in (v) of Assumption 2.4 and we also use the assumption that \( q \leq \frac{1}{3} \kappa_0 \). Consequently, we have

\[
|L''_{\mu_\alpha}(\xi_0)| = \left| \int \frac{1}{(x - \xi_0)^3} d\mu_\alpha(x) \right| \leq (\kappa_0/3)^{-3} \mu_\alpha(\mathbb{R}_+).
\]

Moreover, since \( |\Omega_B(z_0) - \Omega_\beta(z_0)| \leq q \leq \frac{1}{3} \kappa_0 \), we have

\[
|L''_{\mu_\alpha}(\xi_0)\Delta \Omega_2(z_0)^2| \leq 27 \kappa_0^{-3} \mu_\alpha(\mathbb{R}_+) \Delta \Omega_2(z_0)^2 \leq 27 \kappa_0^{-3} \mu_\alpha(\mathbb{R}_+) \|\Delta \Omega(z_0)\|^2 \leq K_3 \|\Delta \Omega(z_0)\|^2.
\]

By (D.9) and (D.10), we readily obtain that

\[
\left| L'_{\mu_\alpha}(\Omega_\beta(z_0)) \Delta \Omega_2(z_0) - \frac{\Delta \Omega_1(z_0)}{z_0} \right| \leq \|r(z_0)\| + K_3 \|\Delta \Omega(z_0)\|^2. \tag{D.11}
\]
Similarly, we have that

\[
\left| L_{\mu,\beta}^\prime (\Omega_{\alpha}(z_0)) \Delta \Omega_1(z_0) - \frac{\Delta \Omega_2(z_0)}{z_0} \right| \leq \|r(z_0)\| + K_3 \|\Delta \Omega(z_0)\|^2. \tag{D.12}
\]

Recall (3.3). We have that

\[
|S_{\alpha,\beta}(z_0)||\Delta \Omega_1(z_0)| = \left| 1 - z_0^2 L_{\mu,\beta}'(\Omega_{\alpha}(z_0))L_{\mu,\beta}'(\Omega_{\beta}(z_0)) \right| |\Delta \Omega_1(z_0)|. \tag{D.13}
\]

Moreover, we note that

\[
|1 - z_0^2 L_{\mu,\beta}'(z_0)L_{\mu,\beta}'(z_0)|
\]

\[
= \left| z_0^2 L_{\mu,\beta}'(\Omega_{\alpha}(z_0)) \left( L_{\mu,\beta}'(\Omega_{\alpha}(z_0)) \Delta \Omega_1(z_0) - \frac{\Delta \Omega_2(z_0)}{z_0} \right) + z_0 \left( L_{\mu,\beta}'(\Omega_{\beta}(z_0)) \Delta \Omega_2(z_0) - \frac{\Delta \Omega_1(z_0)}{z_0} \right) \right|
\]

\[
\leq \left| z_0^2 L_{\mu,\beta}'(\Omega_{\alpha}(z_0)) \right| \left| L_{\mu,\beta}'(\Omega_{\alpha}(z_0)) \Delta \Omega_1(z_0) - \frac{\Delta \Omega_2(z_0)}{z_0} \right| + |z_0| \left| L_{\mu,\beta}'(\Omega_{\beta}(z_0)) \Delta \Omega_2(z_0) - \frac{\Delta \Omega_1(z_0)}{z_0} \right|
\]

\[
\leq (\|z_0^2 L_{\mu,\beta}'(\Omega_{\alpha}(z_0))\| + |z_0|)\|r(z_0)\| + K_3 \|\Delta \Omega(z_0)\|^2,
\]

where we used (D.11) and (D.12) in the first inequality. Together with (D.13), we readily see that

\[
|S_{\alpha,\beta}(z_0)||\Delta \Omega_1(z_0)| \leq (|z_0| + |z_0^2 L_{\mu,\beta}'(\Omega_{\alpha}(z_0))|)\|r(z_0)\| + K_3 \|\Delta \Omega(z_0)\|^2 \leq K_4 \|r(z_0)\| + K_3K_1 \|\Delta \Omega(z_0)\|^2.
\]

Similarly, we have

\[
|S_{\alpha,\beta}(z_0)||\Delta \Omega_2(z_0)| \leq K_4 \|r(z_0)\| + K_3K_4 \|\Delta \Omega(z_0)\|^2.
\]

Consequently, this yields that

\[
\|\Delta \Omega(z_0)\| \leq |\Delta \Omega_1(z_0)| + |\Delta \Omega_2(z_0)| \leq \frac{2K_4}{|S_{\alpha,\beta}(z_0)|} (\|r(z_0)\| + K_3 \|\Delta \Omega(z_0)\|^2)
\]

\[
= \frac{1}{|S_{\alpha,\beta}(z_0)|} \left( \frac{K_1}{2} \|r(z_0)\| + \frac{1}{2K_2} \|\Delta \Omega(z_0)\|^2 \right).
\]

Note that the above equation can be regarded as a quadratic inequality with respect to \(\|\Delta \Omega(z_0)\|\). By an elementary computation, it is easy to see that if the inequality holds then either one of the followings holds

\[
\|\Delta \Omega(z_0)\| \leq K_1 \|r(z_0)\| |S_{\alpha,\beta}(z_0)| \quad \text{or} \quad \|\Delta \Omega(z_0)\| \geq K_2 |S_{\alpha,\beta}(z_0)|.
\tag{D.14}
\]

where we use the trivial bounds

\[
1 - \sqrt{1-x} \leq x, \quad 1 + \sqrt{1-x} \geq 1.
\]

Since the second inequality of (D.14) contradicts the assumption (iii) that \(q \leq \frac{1}{2}K_2S_{\alpha,\beta}(z_0)\). Therefore, we complete the proof. \(\Box\)

### D.2 Proofs of Lemmas 4.6, 4.7 and 5.1

**Proof of Lemma 4.6** Our proof is similar to Lemma S.5.6 of [20] and Lemmas 5.4 and 5.5 of [13]. We start with the first part of the results and first show that each \(\Omega_B^{-1}(B_{\rho}^{-1}(\hat{a}_i)), \pi_a(i) \in S\) is a subset of \(D_{\text{out}}(\omega)\) in (4.16). By (4.7), it is easy to see that \(\Omega_B^{-1}(\zeta) \leq \omega^{-1}\) for all \(\zeta \in \mathbb{C}\) as long as \(\omega\) is sufficiently small. For the lower bound, we claim that for any constant \(C > 0\) and sufficiently small constant \(c_0 < 1\), there exists a constant \(c_1 \equiv c_1(c_0, C)\) such that

\[
\text{Re} \Omega_B^{-1}(\zeta) \geq E_+ + c_1(\text{Re} \zeta - \Omega_B(E_+))^2,
\tag{D.15}
\]

91
for $\Re \zeta \geq \Omega_B(E_+)$, $|\Im \zeta| \leq \tilde{C}(\Re \zeta - \Omega_B(E_+))$ and $|\zeta| \leq \tilde{C}$. By the definition of $\Omega_B^{-1}(B_{\rho^a}(\tilde{a}_i))$ and the definition of $\rho^a_i$, whenever $\zeta \in \Omega_B^{-1}(B_{\rho^a_i}(\tilde{a}_i))$, we have that

$$\Im \zeta \leq c_i(\tilde{a}_i - \Omega_B(E_+)).$$

Moreover, since $\pi_a(i) \in S$, we have that $\Re \zeta \geq \Omega_B(E_+)$. Again using the definition of $\rho^a_i$, we immediately see that

$$\Re \zeta - \Omega_B(E_+) \geq (1 - c_i)(\tilde{a}_i - \Omega_B(E_+)).$$

Together with (D.15), the assumptions $\hat{a}_i - \Omega_B(E_+) \geq N^{-1/3+\tau_1}$ and $\omega < \tau_1/2$, we can conclude the proof when $c_i$’s are sufficiently small. It remains to prove the claim (D.15). The proof follows from an argument similar to equation (S.11) of [20] and a discussion similar to (4.7). We omit the details here. Similarly, we can show that $\Omega_B^{-1}(B_{\rho^a_i}(\Omega_B^{-1}(\tilde{b}_i)))) \subset \mathcal{D}_{out}(\omega)$. This concludes the proof of the first part of the results.

For the second part of the results, it suffices to prove the following:

(i) $\hat{\lambda}_{\pi_a(i)} \in \Omega_B^{-1}(B_{\rho^a_i}(\tilde{a}_i)))$ and $\hat{\lambda}_{\pi_b(\mu)} \in \Omega_B^{-1}(B_{\rho^a_i}(\Omega_B^{-1}(\tilde{b}_i))))$ for all $\pi_a(i) \in S$ and $\pi_b(\mu) \in S$;

(ii) All the other eigenvalues $\hat{\lambda}_j$ satisfy $\hat{\lambda}_j \notin \Omega_B^{-1}(B_{\rho^a_i}(\tilde{a}_i)))$ and $\hat{\lambda}_j \notin \Omega_B^{-1}(B_{\rho^a_i}(\Omega_B^{-1}(\tilde{b}_i))))$ for all $\pi_a(i) \in S$ and $\pi_b(\mu) \in S$.

The proof of (i) and (ii) follows from an analogous discussion to the counterpart in Lemma S.5.6 of [20]. We omit the details here.

**Proof of Lemma 4.7.** The proof is similar to Lemma S.5.7 of [20]. Due to similarity, we only prove (ii). First, the upper bound simply follows from the triangle inequality

$$|\zeta - \hat{a}_i| \leq \rho^a_i + |\hat{a}_i - \tilde{a}_i| \sim \rho^a_i + \delta_{\pi_a(i),\pi_a(j)}^a.$$

Second, we provide a lower bound. When $\pi_a(j) \notin S$, using the fact $|\tilde{a}_i - \tilde{a}_j| \geq 2\rho^a_i$, we have that

$$|\zeta - \hat{a}_j| \geq \rho^a_i + |\tilde{a}_i - \tilde{a}_j|.$$

When $\pi_a(j) \in S$, denote $\delta := |\tilde{a}_i - \tilde{a}_j| - \rho^a_i - \rho^a_j$. In the case when $C_0\delta > |\tilde{a}_i - \tilde{a}_j|$ for some constant $C_0 > 1$, we have that

$$\rho^a_i + \rho^a_j \leq \frac{C_0 - 1}{C_0}|\tilde{a}_i - \tilde{a}_j|.$$

Consequently, we obtain that

$$|\zeta - \hat{a}_j| \geq |\tilde{a}_i - \tilde{a}_j| - \rho^a_i \geq \frac{1}{C_0}|\tilde{a}_i - \tilde{a}_j| \sim \rho^a_i + \delta_{\pi_a(i),\pi_a(j)}^a.$$

In the other case when $C_0\delta \leq |\tilde{a}_i - \tilde{a}_j|$, we have that

$$|\tilde{a}_i - \tilde{a}_j| \leq \frac{C_0}{C_0 - 1}(\rho^a_i + \rho^a_j).$$

To provide a bound regarding $\rho^a_i$ and $\rho^a_j$, we claim that for large enough constant $C_0 > 0$, there exists some constant $\tilde{C} > 1$ such that

$$\frac{1}{C} \rho^a_i \leq \rho^a_j \leq \tilde{C} \rho^a_i.$$  \hspace{1cm} (D.16)

Combining the above bounds, we can prove (4.33). For the justification of (D.16), we can take the verbatim of the proof of equation (S.27) of [20] to finish it. We omit the details here. This concludes our proof. \hspace{1cm} \square
Proof of Lemma 5.1. The proof is similar to Lemma 10.1 of [7]. Recall we have shown that \( \Lambda \prec (N\eta)^{-1} \) holds uniformly in \( z \in D_\tau(\eta_L, \eta_U) \) in (8.38). Suppose that \( \Lambda \prec \hat{\Lambda} \) for some deterministic \( \hat{\Lambda}(z) \) that satisfies

\[
N^\epsilon \left( \frac{1}{N \sqrt{\kappa + \eta \eta}} + \frac{1}{\sqrt{\kappa + \eta \eta}} \frac{1}{(N\eta)^2} \right) \leq \hat{\Lambda}(z) \leq \frac{N^\epsilon}{N\eta}, \tag{D.17}
\]

where \( \epsilon \) is the same as in [7.33]. We point out that such a \( \hat{\Lambda} \) always exists on \( \tilde{D}_\tau \). Since \( \Lambda \prec (N\eta)^{-1} \), by (7.33) and (ii) and (iii) of Proposition 3.1, we conclude that for \( \iota = A, B \), and \( z \in \tilde{D}_\tau \),

\[
|S_{AB}(z)| \Lambda_\iota + T_\iota \Lambda_\iota^2 \prec \sqrt{\left( \frac{\eta}{\sqrt{k + \eta}} + \hat{\Lambda} \right) \left( \sqrt{k + \eta} + \hat{\Lambda} \right)} \frac{1}{N\eta} + \frac{1}{(N\eta)^2} \prec \sqrt{\hat{\Lambda} \sqrt{k + \eta}} + \frac{\sqrt{\eta}}{N\eta} + \frac{1}{(N\eta)^2}, \tag{D.18}
\]

where we used the fact that \( \hat{\Lambda} \prec \frac{N^\epsilon}{N\eta} \leq N^{-\epsilon} \sqrt{k + \eta} \) for all \( z \in \tilde{D}_\tau \). Moreover, again using (iii) of Proposition 3.1, we get

\[
|\Lambda_\iota| \prec \frac{1}{N\eta} \leq N^{-2\epsilon} \sqrt{k + \eta} \sim N^{-2\epsilon} |S_{AB}|. \tag{D.19}
\]

Since \( |T_\iota| \leq C \) for some constant \( C > 0 \) by (iii) of Proposition 3.1, combining (D.19) and (D.18), we obtain that

\[
|\Lambda_\iota| \prec \frac{1}{\sqrt{k + \eta}} \left( \frac{\sqrt{\Lambda} \sqrt{k + \eta}}{N\eta} + \frac{\sqrt{\eta}}{N\eta} + \frac{1}{(N\eta)^2} \right) \leq \frac{1}{N\eta(k + \eta)^{1/4}} \hat{\Lambda}^{1/2} + N^{-\epsilon} \hat{\Lambda} \leq N^{-\epsilon/4} \hat{\Lambda},
\]

where we used (iii) of Proposition 3.1 and (D.17).

From the above argument, we see that we have improved the bound from \( \Lambda \leq \hat{\Lambda} \) to \( \Lambda \leq N^{-\epsilon/4} \hat{\Lambda} \) as long as the lower bound in (D.17) holds. The proof then follows from an iterative improvement.

References

[1] 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.

[2] Z. Bao, X. Ding, J. Wang, and K. Wang. Statistical inference for principal components of spiked covariance matrices. arXiv preprint arXiv 2008.11903, 2020.

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

[4] Z. Bao, L. Erdős, and K. Schnelli. Convergence rate for spectral distribution of addition of random matrices. Advances in Mathematics, 319:251 – 291, 2017.

[5] Z. Bao, L. Erdős, and K. Schnelli. Local law of addition of random matrices on optimal scale. Communications in Mathematical Physics, 349(3):947–990, 2017.

[6] Z. Bao, L. Erdős, and K. Schnelli. Local single ring theorem on optimal scale. Ann. Probab., 47(3):1270–1334, 2019.

[7] 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.

[8] S. T. Belinschi. A note on regularity for free convolutions. Ann. Inst. Henri Poincaré Probab. Stat., 42(5):635–648, 2006.
[9] 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.

[10] F. Benaych-Georges and A. Knowles. Lectures on the local semicircle law for Wigner matrices. *arXiv preprint arXiv 1601.04055*, 2016.

[11] H. Bercovici and D. Voiculescu. Free convolution of measures with unbounded support. *Indiana Univ. Math. J.*, 42(3):733–773, 1993.

[12] A. Bloemendal, L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. Isotropic local laws for sample covariance and generalized Wigner matrices. *Electron. J. Probab.*, 19:53 pp., 2014.

[13] A. Bloemendal, A. Knowles, H.-T. Yau, and J. Yin. On the principal components of sample covariance matrices. *Probability Theory and Related Fields*, 164(1):459–552, 2016.

[14] J. Bun, R. Allez, J. Bouchaud, and M. Potters. Rotational invariant estimator for general noisy matrices. *IEEE Transactions on Information Theory*, 62(12):7475–7490, 2016.

[15] J. Bun, J.-P. Bouchaud, and M. Potters. Cleaning large correlation matrices: Tools from random matrix theory. *Physics Reports*, 666:1 – 109, 2017.

[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] P. Diaconis and M. Shahshahani. The subgroup algorithm for generating uniform random variables. *Probability in the Engineering and Informational Sciences*, 1(1):15–32, 1987.

[18] X. Ding. Spiked sample covariance matrices with possibly multiple bulk components. *Random Matrices: Theory and Applications (in press)*, 2020.

[19] X. Ding and H.-T. Wu. On the spectral property of kernel-based sensor fusion algorithms of high dimensional data. *IEEE Transactions on Information Theory (in press)*, 2020.

[20] X. Ding and F. Yang. Spiked separable covariance matrices and principal components. *The Annals of Statistics (in press)*, 2020.

[21] X. Ding and F. Yang. Tracy-Widom distribution for the edge eigenvalues of Gram type random matrices. *arXiv preprint arXiv 2008.04166*, 2020.

[22] E. Dobriban and S. Liu. Asymptotics for sketching in least squares regression. *Conference on Neural Information Processing Systems (NIPS)*, 2019.

[23] L. Erdős, A. Knowles, and H.-T. Yau. Averaging fluctuations in resolvents of random band matrices. *Ann. Henri Poincaré*, 14(8):1837–1926, 2013.

[24] L. Erdős and H. Yau. *A Dynamical Approach to Random Matrix Theory*. Volume 28 of Courant Lecture Notes. Courant Institute of Mathematical Sciences, New York University and American Mathematical Soc., 2017.

[25] O. P. Ferreira and B. F. Svaiter. Kantorovich’s Theorem on Newton’s Method. *arXiv preprint arXiv 1209.5704*, 2012.

[26] B. Hall. *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*. Graduate Texts in Mathematics. Springer, 2003.

[27] H. C. Ji. Regularity Properties of Free Multiplicative Convolution on the Positive Line. *International Mathematics Research Notices*, 2020. rnaa152.
[28] I. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer New York, 2013.

[29] A. Knowles and J. Yin. The isotropic semicircle law and deformation of Wigner matrices. *Communications on Pure and Applied Mathematics*, 66(11):1663–1749, 2013.

[30] J. Kwak, J. O. Lee, and J. Park. Extremal eigenvalues of sample covariance matrices with general population. *arXiv preprint arXiv 1908.07444*, 2019.

[31] J. Lacotte and M. Pilanci. Effective Dimension Adaptive Sketching Methods for Faster Regularized Least-Squares Optimization. In *Conference on Neural Information Processing Systems (NIPS)*, 2020.

[32] B. Landon and H.-T. Yau. Edge statistics of Dyson Brownian motion. *arXiv preprint arXiv 1712.03881*, 2017.

[33] E. László, H.-T. Yau, and J. Yin. Rigidity of eigenvalues of generalized Wigner matrices. *Advances in Mathematics*, 229(3):1435 – 1515, 2012.

[34] S. Liu and E. Dobriban. Ridge Regression: Structure, Cross-Validation, and Sketching. In *International Conference on Learning Representations (ICLR)*, 2020.

[35] D. Paul and J. W. Silverstein. No eigenvalues outside the support of the limiting empirical spectral distribution of a separable covariance matrix. *Journal of Multivariate Analysis*, 100(1):37 – 57, 2009.

[36] A. Perry, A. S. Wein, A. S. Bandeira, and A. Moitra. Optimality and sub-optimality of PCA I: Spiked random matrix models. *Ann. Statist.*, 46(5):2416–2451, 2018.

[37] V. Vasilchuk. On the law of multiplication of random matrices. *Math. Phys. Anal. Geom.*, 4(1):1–36, 2001.

[38] D. Voiculescu. Multiplication of certain noncommuting random variables. *J. Operator Theory*, 18(2):223–235, 1987.

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

[40] D. Voiculescu, K. Dykema, and A. Nica. *Free Random Variables*. CRM monograph series. American Mathematical Society, 1992.

[41] F. Yang. Edge universality of separable covariance matrices. *Electron. J. Probab.*, 24:57 pp., 2019.

[42] F. Yang, S. Liu, E. Dobriban, and D. P. Woodruff. How to reduce dimension with PCA and random projections? *arXiv preprint arXiv:2005.00511*, 2020.

[43] J. Yao, S. Zheng, and Z. Bai. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015.