The Annals of Applied Probability
2023, Vol. 33, No. 4, 2981–3009
https://doi.org/10.1214/22-AAP1882

LOCAL LAWS FOR MULTIPLICATION OF RANDOM MATRICES

BY XIUCAI DING[1,a] AND HONG CHANG JI[2,b]

[1]*Department of Statistics, University of California, Davis,* [a]*xcading@ucdavis.edu*
[2]*Institute for Science and Technology Austria,* [b]*hongchang.ji@ist.ac.at*

Consider the random matrix model $A^{1/2}UBU^*A^{1/2}$, where A and B are two $N \times N$ deterministic matrices and U is either an $N \times N$ Haar unitary or orthogonal random matrix. It is well known that on the macroscopic scale (*Invent. Math.* **104** (1991) 201–220), the limiting empirical spectral distribution (ESD) of the above model is given by the free multiplicative convolution of the limiting ESDs of A and B, denoted as $\mu_\alpha \boxtimes \mu_\beta$, where μ_α and μ_β are the limiting ESDs of A and B, respectively. In this paper, we study the asymptotic microscopic behavior of the edge eigenvalues and eigenvectors statistics. We prove that both the density of $\mu_A \boxtimes \mu_B$, where μ_A and μ_B are the ESDs of A and B, respectively and the associated subordination functions have a regular behavior near the edges. Moreover, we establish the local laws near the edges on the optimal scale. In particular, we prove that the entries of the resolvent are close to some functionals depending only on the eigenvalues of A, B and the subordination functions with optimal convergence rates. Our proofs and calculations are based on the techniques developed for the additive model $A + UBU^*$ in (*J. Funct. Anal.* **271** (2016) 672–719; *Comm. Math. Phys.* **349** (2017) 947–990; *Adv. Math.* **319** (2017) 251–291; *J. Funct. Anal.* **279** (2020) 108639), and our results can be regarded as the counterparts of (*J. Funct. Anal.* **279** (2020) 108639) for the multiplicative model.

1. Introduction. Large-dimensional random matrices play important roles in high-dimensional statistics. More specifically, given a data matrix Y, studying the eigenvalues and eigenvectors of YY^* and Y^*Y has been known to be an effective approach to analyze the data. There have been many different models for Y depending on the nature of data it represents, and the most fundamental one is the sample covariance matrix [49]. In this context, Y can be written as $Y = A^{1/2}X$, where A is the population covariance matrix and X contains i.i.d. centered random variables. An extension of the sample covariance matrix is the separable covariance matrix [15, 22, 23, 41, 47], where $Y = A^{1/2}XB^{1/2}$ with another positive definite matrix B.

Even though the assumption that X has i.i.d. entries is popular and useful in the literature, its applications are limited to data composed of linear functions of independent samples. An important example that such an assumption cannot cover is the Haar distributed random matrix, which has been used in the literature of statistical learning theory [24, 35, 38, 48]. More specifically, we consider $X = U$ to be either an $N \times N$ random Haar unitary or orthogonal matrix so that

$$Y = A^{1/2}UB^{1/2}. \tag{1.1}$$

In other words, we study a general class of separable random matrices beyond the i.i.d. assumption; indeed, the model (1.1) covers the case where X consists of i.i.d. Gaussian random variables due to invariance. We mention that the data matrix (1.1) has appeared in the study

Received August 2021; revised May 2022.
MSC2020 subject classifications. Primary 46L54, 60B20; secondary 15B52.
Key words and phrases. Random matrices, free multiplicative convolution, subordination functions, edge statistics.

of high-dimensional data analysis, for instance, data acquisition [19], matrix denoising [14, 15] and random sketching [24, 48].

The empirical spectral distribution (ESD) of YY^* in (1.1) has been studied in the literature of free probability theory. Denote

$$H = AUBU^*, \tag{1.2}$$

which has the same eigenvalues with YY^*. In the influential work [45], Voiculescu studied the limiting spectral distribution of the eigenvalues of H and showed that it was given by the free multiplicative convolution of μ_α and μ_β, denoted as $\mu_\alpha \boxtimes \mu_\beta$, where μ_α and μ_β are the limiting ESDs of A and B, respectively; see Definition 2.7 for a precise statement. More recently, in [31], the author investigated the behavior of $\mu_\alpha \boxtimes \mu_\beta$ by analyzing a system of deterministic equations, known as subordination equations, that defines the free convolution; see equation (2.14) for details. They also proved that under certain regularity assumptions, the density of $\mu_\alpha \boxtimes \mu_\beta$ had a regular square root behavior near the edges of its support.

However, on the microscopic level, the singular value and vector statistics of Y, as well as the local laws, have not been established so far. The aim of this paper is to fill this gap near the regular edges. Before proceeding to our main focus, we pause to discuss the additive model, that is, $A + UBU^*$. The ESD of the additive model converges to the free additive convolution of μ_α and μ_β, denoted as $\mu_\alpha \boxplus \mu_\beta$ [45]. More recently, the local laws as well as eigenvalues and eigenvectors statistics have been extensively studied in the series of papers [2–4, 6, 7]. Our arguments are strongly inspired by these works and our results can be regarded as multiplicative counterparts of [7]. In what follows, we highlight and summarize the results and techniques of the additive model [2–4, 7] in Section 1.1. Then we explain how we adapt their approaches with some modifications to obtain the results for the multiplicative model (1.2) in Section 1.2.

1.1. *Local laws for addition of random matrices.* In this subsection, we review the results and techniques for the addition of random matrices $A + UBU^*$ in the series of papers [2–4, 7].

In [2–4], the authors studied the local laws in regular bulk spectrum of the free additive convolution. Chronologically, in [2], the authors proved that the system of the subordination equations, defining the free additive convolution, was stable away from the edges of the support and singularities. In particular, on one hand, they showed that the system was stable and the imaginary parts of the subordination functions were bounded below in the regular bulk; on other hand, they proved a local stability result of the free additive convolution. Based on [2], in [3], they proved that the local laws held in the bulk of the spectrum down to the optimal scale $N^{-1+\gamma}$, for any $\gamma > 0$, which improved a result obtained in [2]. Particularly, they proved a version of averaged local law that the ESD of $A + UBU^*$ concentrated around $\mu_A \boxplus \mu_B$ where μ_A and μ_B denote the ESDs of A and B, respectively. They also proved the entrywise local law that the every entry of the resolvent $G := (A + UBU^* - z)^{-1}$, $z = E + \mathrm{i}\eta \in \mathbb{C}_+$, was well estimated at deterministic functions of z. As a byproduct, they showed that the bulk eigenvectors were completely delocalized. Later on, in [4], the authors obtained the optimal convergence rate $(N\eta)^{-1}$ in the bulk for the local laws which improved the result of [3] where the convergence speed was shown to be of smaller order than $(N\eta)^{-1/2}$.

We highlight several important technical components and insights of the aforementioned three works. Since [4] established the local laws down to the optimal scale with optimal precision which refined the results of [2, 3], we focus our discussion on [4]. The core is to explore the system of subordination functions globally and locally. First, since the additive model lacks the independence of matrix elements, they employed a partial randomness decomposition (see (3.1) in the present paper) of the Haar measure which enabled them to take

partial expectations of the entries of the resolvent. Second, to connect the resolvent with the subordination functions, they used the approximate subordination functions which depend only on the resolvent of $A + UBU^*$. In particular, their choices for these approximates can be considered as a random version of those used in [32, 40]; see equation (3.18) of [4]. With such choices, they were able to work on a new system of self-consistent equations. Surprisingly, it suffices to monitor only two auxiliary quantities to analyze the system. With the aid of the local stability results of [2], they connected the partial expectations of the entries of the resolvent with the subordination functions. Third, they proposed a novel strategy to handle the fluctuation averaging mechanism for Haar random matrices. More specifically, instead of working directly with $N^{-1}\sum_i G_{ii}$, they first considered generic averages of an auxiliary quantity which was a carefully chosen linear combination of G_{ii} and $(UBU^*G)_{ii}$. Such a particular choice made the leading order terms within its average cancel algebraically, and the auxiliary quantity can be passed to G_{ii} by taking different weights for this average. Finally, to streamline the calculation, instead of directly computing high moments of the essential auxiliary quantities, they used the so-called recursive moment estimates, in which high-moments were estimated in terms of the lower moments with the aid of integration by parts.

Armed with the above techniques and results, in [7], they were able to investigate the local laws near the regular edges in the sense that $\mu_A \boxplus \mu_B$ had a regular behavior near the edges. More specifically, they presented the local laws near the edges on the optimal scale with optimal precision. Based on these results, they were able to prove the edge eigenvalue rigidity and edge eigenvector delocalization. On the technical level, they used and generalized the strategies and inputs of [2–4] as summarized in the previous paragraph. Since the eigenvalues around the edges are sparse and fluctuate more, in order to guarantee the regular behavior of $\mu_A \boxplus \mu_B$, they first established the square root decay of their limiting counterpart $\mu_\alpha \boxplus \mu_\beta$. Under suitable assumptions on the Lévy distances, with the local stability, the measure $\mu_A \boxplus \mu_B$ inherits the regularity up to the optimal scale. In addition, the probabilistic part of [4] is not sufficient around the edges as the subordination functions become unstable and the improvement from fluctuation averaging in [4] is suboptimal. In order to compensate this instability, they established a very accurate estimate on the approximation error. To achieve this goal, they carefully identified a new pair of auxiliary quantities; see equations (4.14) and (4.15) of [7]. In particular, one of the auxiliary quantities in [7] has an additional counter term compared to the one used in [4]. We mention that [7] required the assumption that at least one of the Stieltjes transforms of μ_α and μ_β was bounded from above. This assumption can be removed using their recent results in [6].

In summary, using addition of random matrices as an example, the authors in [2–4, 7] have developed a general framework and powerful techniques to study the local laws of random matrix models where the main source of randomness is the Haar matrix. Since multiplication of random matrices is another typical example using Haar matrix, it is natural to study the multiplication of random matrices using the techniques developed for the additive model. This will be discussed in the next subsection, Section 1.2.

1.2. *From addition to multiplication*: *An overview of our results*. In this subsection, we explain how to adapt the techniques of the additive model [2–4, 7] as summarized in Section 1.1 to obtain the results for the multiplicative model (1.2). The main purpose of this paper is to present a comprehensive edge local law on the optimal scale and with optimal convergence rates for the multiplicative model, which is the counterpart of [7]. In what follows, we give an overview of our results and explain how to handle the multiplicative model adapting the techniques of the additive model in [4, 7].

The first part of our results concerns the regularity of $\mu_A \boxtimes \mu_B$ and the subordination functions. More specifically, in Proposition 2.11 below, we establish the stability properties

of the subordination functions near the regular edges and provide some crucial estimates. Here we point out that, instead of using the conventional η-transform [10, 44] to define the subordination functions and free multiplicative convolution, we use a simple conjugate of it known as M-transform (cf. Definition 2.1) [16, 31]. One technical advantage of using M-transform is that it makes the similarity between the additive and multiplicative models more evident, which enables us to adapt the techniques of [4, 7] more directly and easily. The proof of Proposition 2.11 follows from its counterpart for the additive model in [7] (see Proposition 3.1 therein) which can be split into two steps. In the first step, the results are proved for the limiting measures μ_α and μ_β under some regularity assumptions (cf. Assumption 2.2). In the second step, assuming that μ_A, μ_B and μ_α, μ_β are close enough (cf. Assumption 2.4), the statements can be carried over to the measures μ_A and μ_B. As mentioned earlier, in [7], the authors proved analogous results for the additive model assuming that at least one of the Stieltjes transforms of μ_α and μ_β was bounded from above (see (iii) of Assumption 2.1 in [7]), which could be removed using their recent results in [6]. For our multiplicative model, since the analog of [6] has been established by the second author in [31], we will not need this condition in our Assumption 2.2.

The second part of our results focuses on establishing the optimal edge local laws on the optimal scale for the multiplicative model. In Theorem 2.13 below, we provide accurate estimates for the entries of the resolvent and also prove the averaged local law. The convergence rates are optimal up to some N^ϵ factor. As two consequences, we prove the rigidity of the edge eigenvalues in Theorem 2.15 and the complete delocalization of the edge eigenvectors in Theorem 2.16. On the technical level, the proof of Theorem 2.13 follows closely from its counterpart for the additive model [7] (see Theorem 2.5 therein) as summarized in Section 1.1. In what follows, we highlight the key ingredients on the adaption of their arguments. Thanks to the M-transform, the approximate subordination functions for the multiplicative model (cf. Definition 3.1) can be easily identified. To control the errors between the subordination functions and their approximates, we first explore some hidden relations. For instance, in (3.7) and (3.10) we represent the error in terms of the resolvents. This enables us to find the key auxiliary quantities to work with. Then we use integration by parts to start the recursive estimates to obtain bounds for high moments of these essential quantities. In order to establish the optimal convergence rates, as mentioned in [7], the weights in the fluctuation averaging mechanism needed to be properly chosen. In our case, these weights (cf. equations (B.31) and (B.32) in our supplementary file [18]) can be constructed using the hidden identities obtained earlier. Finally, we point that due to the structural difference between the additive and multiplicative models, many errors in our model need more careful treatment. For example, in the fluctuation averaging mechanism, our error terms e_{i1} in (4.24) and e_{i2} in (4.49) will generate some $\mathrm{O}_\prec(N^{-1/2})$ terms. The weighted summations of these terms will be canceled out algebraically after we explore some hidden identities; see (B.14)–(B.17) and the associated discussion in [18] for more details.

As mentioned in [7], the results of the addition of random matrices demonstrate that the Haar randomness in the additive model leads to an analogous behavior to the Wigner matrices [28] in the sense of strong concentration of the eigenvalues and eigenvectors. In the same spirit, the Haar randomness in our multiplicative model (1.1), results in a similar behavior as the separable covariance matrices as in [22, 47]. Finally, we mention that the arguments of the current paper can be carried out to study the bulk eigenvalues and eigenvectors as in [2–4] which deals with additive model. The results obtained here can also be used to study other models and statistics, for example, the deformed invariant model [11] and the Tracy–Widom distribution for the edge eigenvalues. We will pursue these topics in future works.

The rest of the paper is organized as follows. In Section 2, we introduce the necessary notation and state the main results. In Section 3, we present a structural summary of our

proof. In Section 4, we prove a subordination property for the resolvent entries. The proof of fluctuation averaging lemmas, along with some auxiliary lemmas and technical proofs, are collected in our supplementary file [18].

Conventions. For $M, N \in \mathbb{N}$, we denote $\{k \in \mathbb{N} : M \le k \le N\}$ by $[\![M, N]\!]$. For $N \in \mathbb{N}$ and $i \in [\![1, N]\!]$, we denote by $\mathbf{e}_i^{(N)} \in \mathbb{R}^N$ with $(\mathbf{e}_i)_j = \delta_{ij}$. We often omit the superscript N to write $\mathbf{e}_i^{(N)} \equiv \mathbf{e}_i$. We use I for the identity matrix of any dimension. For an N-dimensional real or complex random vector $\boldsymbol{g} = (g_1, \dots, g_N)$, we write $\boldsymbol{g} \sim \mathcal{N}_{\mathbb{R}}(0, \sigma^2 I_N)$ if $g_1, \dots, g_N$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables, and we write $\boldsymbol{g} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2 I_N)$ if $g_1, \dots, g_N$ are i.i.d. $\mathcal{N}_{\mathbb{C}}(0, \sigma^2)$ variables, where $g_i \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$ means that $\operatorname{Re} g_i$ and $\operatorname{Im} g_i$ are independent $\mathcal{N}(0, \frac{\sigma^2}{2})$ random variables. For any matrix A, we denote its operator norm by $\|A\|$ and for a vector $\boldsymbol{v}$, we use $\|\boldsymbol{v}\|$ for its ℓ_2 norm.

2. Main results.

2.1. *Notation and assumptions.* For any $N \times N$ matrix W, we denote its normalized trace by $\operatorname{tr} W$, that is,

$$\operatorname{tr} W = \frac{1}{N} \sum_{i=1}^{N} W_{ii}. \tag{2.1}$$

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.

Consider two $N \times N$ real deterministic positive definite matrices

$$A \equiv A_N = \operatorname{diag}(a_1, \dots, a_N), \qquad B \equiv B_N = \operatorname{diag}(b_1, \dots, b_N),$$

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

$$\begin{aligned} &H := AUBU^*, \qquad \mathcal{H} := U^* AUB, \qquad \widetilde{H} := A^{1/2} \widetilde{B} A^{1/2} \quad \text{and} \\ &\widetilde{\mathcal{H}} := B^{1/2} \widetilde{A} B^{1/2}. \end{aligned} \tag{2.2}$$

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

Since H, $\mathcal{H}$, $\widetilde{H}$ and $\widetilde{\mathcal{H}}$ have the same eigenvalues, in the sequel, we denote the eigenvalues of all of them as $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_N$ without causing any confusion. Further, we define the ESDs of the above matrices by

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

For $z \in \mathbb{C}_+ := \{z \in \mathbb{C} : \operatorname{Im} z > 0\}$, we define the resolvent of H as

$$G(z) := (H - zI)^{-1}. \tag{2.3}$$

Similarly, the resolvents of $\mathcal{H}$, $\widetilde{H}$ and $\widetilde{\mathcal{H}}$ are defined as $\mathcal{G}(z)$, $\widetilde{G}(z)$ and $\widetilde{\mathcal{G}}(z)$, respectively. In the rest of the paper, we usually omit the dependence on 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 μ defined on $\mathbb{R}_+$, its *Stieltjes transform* m_μ is defined as

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

Moreover, we define the M-transform M_μ and L-transform L_μ on $\mathbb{C}\setminus\mathbb{R}_+$ as

$$M_\mu(z) := 1 - \left(\int \frac{x}{x-z}\,\mathrm{d}\mu(x)\right)^{-1} = \frac{z m_\mu(z)}{1+z m_\mu(z)}, \qquad L_\mu(z) := \frac{M_\mu(z)}{z}. \tag{2.4}$$

Let $m_H(z)$ be the Stieltjes transform of the ESD of H. Since H, $\mathcal{H}$, $\widetilde{H}$ and $\widetilde{\mathcal{H}}$ are similar to each other, we have that $m_H(z) = \operatorname{tr} G = \operatorname{tr}\mathcal{G} = \operatorname{tr}\widetilde{G} = \operatorname{tr}\widetilde{\mathcal{G}}$. Moreover, we have

$$G_{ij} = \sqrt{a_i/a_j}\,\widetilde{G}_{ij}, \qquad \mathcal{G}_{ij} = \sqrt{b_j/b_i}\,\widetilde{\mathcal{G}}_{ij}. \tag{2.5}$$

With the above preparation, we introduce the main assumptions. Analogous to [7], throughout the paper, we assume that μ_A and μ_B converge to some N-independent absolutely continuous probability measures μ_α and μ_β. We start with stating the assumptions on μ_α and μ_β, which is an analog of the additive model as in [7], Assumption 2.1.

ASSUMPTION 2.2. Suppose the following assumptions hold true:

(i) μ_α and μ_β have densities ρ_α and ρ_β, respectively. For the ease of discussion, we assume that both of them have means 1.

(ii) Both ρ_α and ρ_β have single nonempty intervals as supports, denoted as $[E_-^\alpha, E_+^\alpha]$ and $[E_-^\beta, E_+^\beta]$, respectively. Here E_-^α, E_+^α, E_-^β, and E_+^β are all positive numbers. Moreover, both of the density functions are strictly positive in the interior of their supports.

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

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

$$C^{-1} \le \frac{\rho_\beta(x)}{(x-E_-^\beta)^{t_-^\beta}(E_+^\beta - x)^{t_+^\beta}} \le C \quad \forall x \in \big[E_-^\beta, E_+^\beta\big].$$

REMARK 2.3. First, the assumption that both μ_α and μ_β have means 1 in (i) is introduced for technical simplicity and can be removed easily; see Remark 3.2 of [31]. Second, the assumption (iii) is introduced to guarantee the square root behavior near the edges of the free multiplicative convolution of μ_α and μ_β. When this condition fails, the behavior of $\mu_\alpha \boxtimes \mu_\beta$ near the edge can be very different from our current discussion; see [34] for more details. Third, as we are only interested in the edge statistics near the upper edge in Section 2.3, the assumptions (ii) and (iii) can be relaxed by only imposing conditions on E_+^α and E_+^β. We keep the current form involving E_-^α and E_-^β since our results also hold near the lower edge with minor modification. Finally, for the additive model, the counterpart is Assumption 2.1 of [7]. It requires that at least one of the Stieltjes transforms of μ_α and μ_β is bounded from above (see (iii) of Assumption 2.1 in [7]), which could be removed using their recent results in [6]. We will no longer need this condition since the counterpart of [6] for our model has been established in [31].

The following Assumption 2.4 ensures that μ_A and μ_B are close to μ_α and μ_β, respectively. Specifically, it demonstrates that the convergence rates of μ_A and μ_B to μ_α and μ_β are bounded by an order of N^{-1}, so that their fluctuations do not dominate that of μ_H. Its counterpart for the additive model is [7], Assumption 2.2.

ASSUMPTION 2.4. Suppose the following hold true when N is sufficiently large:

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

$$\boldsymbol{d} := \mathcal{L}(\mu_\alpha, \mu_A) + \mathcal{L}(\mu_\beta, \mu_B) \le N^{-1+\epsilon}.$$

(v) For the supports of μ_A and μ_B, we have that for any constant $\delta > 0$

$$\operatorname{supp}\mu_A \subset \big[E_-^\alpha - \delta, E_+^\alpha + \delta\big], \qquad \operatorname{supp}\mu_B \subset \big[E_-^\beta - \delta, E_+^\beta + \delta\big].$$

REMARK 2.5. We remark that we will consistently use ϵ as a generic sufficiently small constant whose value may change from one line to the next. The assumption (v) assures that both of the upper edges of μ_A and μ_B are bounded.

As proved by Voiculescu in [44, 45], under Assumptions 2.2 and 2.4, μ_H converges weakly to the free multiplicative convolution of μ_α and μ_β, denoted as $\mu_\alpha \boxtimes \mu_\beta$.

LEMMA 2.6 (Proposition 2.5 of [31]). *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*

$$\arg\Omega_\alpha(z) \ge \arg z \quad \textit{and} \quad \arg\Omega_\beta(z) \ge \arg z. \tag{2.6}$$

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

$$\lim_{z\searrow-\infty} \Omega_\alpha(z) = \lim_{z\searrow-\infty} \Omega_\beta(z) = -\infty. \tag{2.7}$$

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

$$zM_{\mu_\alpha}\big(\Omega_\beta(z)\big) = zM_{\mu_\beta}\big(\Omega_\alpha(z)\big) = \Omega_\alpha(z)\Omega_\beta(z). \tag{2.8}$$

The analytic functions Ω_α and Ω_β are referred to as the subordination functions. We remark that the same functions as well as M-transforms also appeared in [16], called Z and K-functions, respectively. Similarly, we can denote Ω_A and Ω_B by replacing (α, β) with (A, B). With the aid of Lemma 2.6, we can 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

$$M(z) := M_{\mu_\alpha}\big(\Omega_\beta(z)\big) = M_{\mu_\beta}\big(\Omega_\alpha(z)\big). \tag{2.9}$$

Then the free multiplicative convolution of μ_α and μ_β is defined as the unique probability measure μ, denoted as $\mu \equiv \mu_\alpha \boxtimes \mu_\beta$ such that (2.9) holds for all $z \in \mathbb{C}\setminus\mathbb{R}_+$. In other words, $M(z) \equiv M_{\mu_\alpha\boxtimes\mu_\beta}(z)$ is the M-transform of $\mu_\alpha \boxtimes \mu_\beta$. Furthermore, we define $\mu_A \boxtimes \mu_B$ so that $M_{\mu_A}(\Omega_B(z)) = M_{\mu_B}(\Omega_A(z)) = M_{\mu_A\boxtimes\mu_B}(z)$ holds for all $z \in \mathbb{C}\setminus\mathbb{R}_+$.

Note that a consequence of (2.8) and the definition of $M_\mu(z)$ is the following identity:

$$\int \frac{x}{x-z}\,\mathrm{d}(\mu_\alpha \boxtimes \mu_\beta)(x) = \Omega_\beta(z)m_{\mu_\alpha}\big(\Omega_\beta(z)\big) + 1 = \int \frac{x}{x-\Omega_\beta(z)}\,\mathrm{d}\mu_\alpha(x). \tag{2.10}$$

REMARK 2.8. Since all of μ_α, μ_β, μ_A and μ_B are compactly supported on $(0,\infty)$, similar results hold for $\mu_\alpha \boxtimes \mu_\beta$ and $\mu_A \boxtimes \mu_B$. Specifically, according to [46], Remark 3.6.2(iii), we have

$$\operatorname{supp}\mu_\alpha \boxtimes \mu_\beta \subset \big[E_-^\alpha E_-^\beta, E_+^\alpha E_+^\beta\big], \qquad \operatorname{supp}\mu_A \boxtimes \mu_B \subset [a_N b_N, a_1 b_1]. \tag{2.11}$$

In fact, we can conclude from [31], Theorem 3.1, that, if (i) and (ii) of Assumption 2.2 hold, $\mu_\alpha \boxtimes \mu_\beta$ is absolutely continuous and supported on a single nonempty compact interval on $(0, \infty)$, denoted as $[E_-, E_+]$, that is,

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

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

$$\rho(x) \sim \sqrt{E_+ - x}, \quad x \in [E_+ - \tau, E_+]. \tag{2.13}$$

Furthermore, as we will see in Lemma A.3 of our supplement [18], the subordination functions Ω_α and Ω_β also have square root behaviors near the edges. The regularity behavior is assured by the fact that the subordination functions Ω_α and Ω_β are well separated from the supports of μ_β and μ_α, respectively; see (ii) of Lemma A.1 in [18]. In fact, from the proof of Proposition 5.6 of [31], we see that the assumption (iii) of Assumption 2.2 implies this stability condition.

REMARK 2.9. It is known from [9, 31] that Assumption 2.2 ensures that the subordination functions $\Omega_\alpha|_{\mathbb{C}_+}$ and $\Omega_\beta|_{\mathbb{C}_+}$ can be extended continuously to the real line. Throughout the paper, we will write $\Omega_\alpha(x)$ or $\Omega_\beta(x)$ for $x \in \mathbb{R}$ to denote 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. *Properties of subordination functions.* In this subsection, we state the results regarding the local properties of the subordination functions and related quantities near the regular edge. These results will be used in the proof of the local laws. We first introduce some notation. Note that the system of subordination equations (2.8) can be rewritten as

$$\Phi_{\alpha\beta}\big(\Omega_\alpha(z), \Omega_\beta(z), z\big) = 0, \tag{2.14}$$

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

$$\Phi_\alpha(\omega_1, \omega_2, z) := \frac{M_{\mu_\alpha}(\omega_2)}{\omega_2} - \frac{\omega_1}{z}, \qquad \Phi_\beta(\omega_1, \omega_2, z) := \frac{M_{\mu_\beta}(\omega_1)}{\omega_1} - \frac{\omega_2}{z}. \tag{2.15}$$

Here $\Phi_{\alpha\beta}$ should be regarded 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 (2.14). Recall (2.4). Denote

$$\mathcal{S}_{\alpha\beta}(z) := z^2 L'_{\mu_\beta}\big(\Omega_\alpha(z)\big) L'_{\mu_\alpha}\big(\Omega_\beta(z)\big) - 1, \tag{2.16}$$

$$\begin{aligned} \mathcal{T}_\alpha(z) &:= \frac{1}{2}\big[z L''_{\mu_\beta}\big(\Omega_\alpha(z)\big) L'_{\mu_\alpha}\big(\Omega_\beta(z)\big) + \big(z L'_{\mu_\beta}\big(\Omega_\alpha(z)\big)\big)^2 L''_{\mu_\alpha}\big(\Omega_\beta(z)\big)\big], \\ \mathcal{T}_\beta(z) &:= \frac{1}{2}\big[z L''_{\mu_\alpha}\big(\Omega_\beta(z)\big) L'_{\mu_\beta}\big(\Omega_\alpha(z)\big) + \big(z L'_{\mu_\alpha}\big(\Omega_\beta(z)\big)\big)^2 L''_{\mu_\beta}\big(\Omega_\alpha(z)\big)\big]. \end{aligned} \tag{2.17}$$

By replacing the pair (α, β) with (A, B), we can define Φ_{AB}, $\mathcal{S}_{AB}$, $\mathcal{T}_A$ and $\mathcal{T}_B$ analogously. We remark that analogous quantities have been defined and used for the additive model in [7]; see equation (3.1) therein.

REMARK 2.10. We provide a few remarks on the usefulness for the above quantities. First, the edges $E_\pm$ of $\mu_\alpha \boxtimes \mu_\beta$ can be completely characterized by the equation $\mathcal{S}_{\alpha\beta}(E\pm) = 0$; see [31], Section 5, for mode details. Second, the above quantities are closely connected

with the subordination equation system (2.14). Let D be the differential operator with respect to ω_1 and ω_2. Then we find that the first derivative of $\Phi_{\alpha\beta}$ is given by

$$\mathrm{D}\Phi_{\alpha\beta}(\omega_1,\omega_2,z) := \begin{pmatrix} -z^{-1} & L'_{\mu_\alpha}(\omega_2) \\ L'_{\mu_\beta}(\omega_1) & -z^{-1} \end{pmatrix}. \tag{2.18}$$

Moreover, its determinant is equal to $-z^{-2}\mathcal{S}_{\alpha\beta}(z)$ 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

$$\mathcal{T}_\alpha(z) = z\left[\frac{\partial}{\partial\omega_1}\det \mathrm{D}\Phi_{\alpha\beta}\big(\omega_1, zL_{\mu_\beta}(\omega_1), z\big)\right]_{\omega_1=\Omega_\alpha(z)},$$

$$\mathcal{T}_\beta(z) = z\left[\frac{\partial}{\partial\omega_2}\det \mathrm{D}\Phi_{\alpha\beta}\big(zL_{\mu_\alpha}(\omega_2), \omega_2, z\big)\right]_{\omega_2=\Omega_\beta(z)}.$$

As will be seen later in the proof, we need to show that Ω_α and Ω_β are close to Ω_A and Ω_B, respectively. The arguments are based on the stability analysis of Φ_{AB}, which require sharp estimates of the above quantities.

We collect the key properties of the subordination functions in Proposition 2.11. It is the counterpart of Proposition 3.1 of [7] which concerns the additive model. For the ease of statements, we only provide the results near the upper edge E_+ defined in (2.12). Similar results hold for the lower edge E_-. For $z=E+\mathrm{i}\eta\in\mathbb{C}_+$, denote

$$\kappa \equiv \kappa(z) := |E-E_+|. \tag{2.19}$$

For some given constants $0\le a\le b$ and $0<\tau<\min\{\frac{E_+-E_-}{2},1\}$, we define the following set of spectral parameters by

$$\mathcal{D}_\tau(a,b) := \big\{z=E+\mathrm{i}\eta\in\mathbb{C}_+ : E_+-\tau\le E\le\tau^{-1}, a\le\eta\le b\big\}. \tag{2.20}$$

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

$$\eta_L \equiv \eta_L(\gamma) := N^{-1+\gamma}, \tag{2.21}$$

and let $\eta_U>1$ be a large N-independent constant.

PROPOSITION 2.11. *Suppose Assumptions 2.2 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*

$$\min_i|a_i-\Omega_B(z)|\ge C^{-1}, \qquad \min_i|b_i-\Omega_A(z)|\ge C^{-1},$$

$$C^{-1}\le|\Omega_A(z)|\le C, \qquad C^{-1}\le|\Omega_B(z)|\le C,$$

uniformly in $z\in\mathcal{D}_\tau(\eta_L,\eta_U)$.

(ii) *For all $z\in\mathcal{D}_\tau(\eta_L,\eta_U)$, we have*

$$\operatorname{Im} m_{\mu_A\boxtimes\mu_B}(z)\sim\begin{cases}\sqrt{\kappa+\eta} & \text{if } E\in\operatorname{supp}\mu_A\boxtimes\mu_B,\\ \dfrac{\eta}{\sqrt{\kappa+\eta}} & \text{if } E\notin\operatorname{supp}\mu_A\boxtimes\mu_B.\end{cases}$$

(iii) *For all $z\in\mathcal{D}_\tau(\eta_L,\eta_U)$, we have the following bounds for $\mathcal{S}_{AB}$, $\mathcal{T}_A$, and $\mathcal{T}_B$,*

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

Furthermore, if $|z-E_+|\le\delta$ for sufficiently small constant $\delta>0$, we also have the lower bounds for $\mathcal{T}_A$ and $\mathcal{T}_B$ such that for some small constant $c>0$

$$|\mathcal{T}_A(z)|\ge c, \qquad |\mathcal{T}_B(z)|\ge c.$$

(iv) *For the derivatives of* Ω_A, Ω_B *and* $\mathcal{S}_{AB}$, *we have*

$$|\Omega_A'(z)| \le C\frac{1}{\sqrt{\kappa+\eta}}, \qquad |\Omega_B'(z)| \le C\frac{1}{\sqrt{\kappa+\eta}}, \qquad |S_{AB}'(z)| \le C\frac{1}{\sqrt{\kappa+\eta}},$$

uniformly in $z \in \mathcal{D}_\tau(\eta_L, \eta_U)$.

Proposition 2.11 will be proved in Section A of our supplement [18]. First, the first equation in (i) states that the subordination functions are well separated from the supports of μ_A and μ_B. This regularity further implies the square root behavior of the subordination functions; see Lemmas A.3 and A.7 of [18]. The second equation in (i) shows that the subordination functions are bounded from both below and above. Second, (ii) offers a standard estimate for the Stieltjes transform, which follows from the square root behavior of $\mu_A \boxtimes \mu_B$. Third, (iii) and (iv) prepare some estimates for the related quantities. All these will be used to prove the closeness between Ω_α, Ω_β and Ω_A, Ω_B; see Section A.3 of [18] for more details.

2.3. *Local laws for free multiplication of random matrices.* In this subsection, we state the results of the local laws. We will need the notion of *stochastic domination*. It was first introduced in [26] and subsequently used in many works on random matrix theory. It simplifies the presentation of the results by systematizing statements of the form "X_N is bounded by Y_N with high probability up to a small power of N".

DEFINITION 2.12. 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 $X_N = \mathrm{O}_\prec(Y_N)$, if for all (small) $\epsilon > 0$ and (large) $D > 0$, we have

$$\mathbb{P}\big[|X_N| \ge N^\epsilon |Y_N|\big] \le N^{-D},$$

for sufficiently large $N \ge N_0(\epsilon, D)$. 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 Ξ holds with high probability if for any constant $D > 0$, $\mathbb{P}(\Xi) \ge 1 - N^{-D}$ for large enough N.

The following theorem establishes the local laws for the matrices H, $\widetilde{H}$, $\mathcal{H}$ and $\widetilde{\mathcal{H}}$ near the upper edge E_+. Analogous results can be obtained for the lower edge E_-. It can be regarded as the counterpart of [7], Theorem 2.5.

THEOREM 2.13. *Suppose Assumptions 2.2 and 2.4 hold. Let* τ *and* γ *be fixed small positive constants. Given any deterministic vector* $\boldsymbol{v} = (v_1, \dots, v_N) \in \mathbb{C}$ *such that* $\|\boldsymbol{v}\|_\infty \le 1$, *the following hold true*:

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

$$\left|\frac{1}{N}\sum_{i=1}^N v_i\left(zG_{ii}(z) + 1 - \frac{a_i}{a_i - \Omega_B(z)}\right)\right| \prec \frac{1}{N\eta}, \tag{2.22}$$

uniformly in $z \in \mathcal{D}_\tau(\eta_L, \eta_U)$ *with* η_L *in* (2.21) *and any fixed constant* η_U. *Particularly*,

$$|m_H(z) - m_{\mu_A\boxtimes\mu_B}(z)| \prec \frac{1}{N\eta}. \tag{2.23}$$

Moreover, we have the following entrywise local law:

$$\max_{i,j}\left|G_{ij}(z) - \delta_{ij}\frac{\Omega_B(z)}{z(\Omega_B(z) - a_i)}\right| \prec \frac{1}{\sqrt{N\eta}}. \tag{2.24}$$

Similar results hold true by replacing H *and* $G(z)$ *with* $\widetilde{H}$ *and* $\widetilde{G}(z)$, *respectively.*

(2) *For the matrix $\mathcal{H}$ and its resolvent $\mathcal{G}(z)$, we have*

$$\left|\frac{1}{N}\sum_{i=1}^{N} v_i\left(z\mathcal{G}_{ii}(z)+1-\frac{b_i}{b_i-\Omega_A(z)}\right)\right| \prec \frac{1}{N\eta}, \tag{2.25}$$

and

$$\left|m_{\mathcal{H}}(z)-m_{\mu_A\boxtimes\mu_B}(z)\right| \prec \frac{1}{N\eta},$$

uniformly in $z\in\mathcal{D}_\tau(\eta_L,\eta_U)$. Moreover, for the entrywise local law, we have

$$\max_{i,j}\left|\mathcal{G}_{ij}(z)-\delta_{ij}\frac{\Omega_A(z)}{z(\Omega_A(z)-b_i)}\right| \prec \frac{1}{\sqrt{N\eta}}. \tag{2.26}$$

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

REMARK 2.14. We provide a few remarks for Theorem 2.13. First, since the goal of [7] is to establish the spectral rigidity for the additive model, they only need the averaged local laws so the entrywise local laws are not presented explicitly there. However, it is easy to check that such entrywise laws also hold for the additive model following their proofs. In fact, the entrywise local laws are stated explicitly for the additive model in the regular bulk in their work [3] (see Theorem 2.5 therein). Second, it is not hard to check that for the entrywise local laws, the convergence rates can be replaced by

$$\sqrt{\frac{\operatorname{Im} m_{\mu_A\boxtimes\mu_B}(z)}{N\eta}}+\frac{1}{N\eta},$$

which matches the typical forms of the bounds of local laws in the random matrix theory literature; see the monograph [28]. We keep the current form to highlight the similarities between our multiplicative model and the additive model in [7]. Third, in [7], the authors also state the averaged local law far away from the edges such that the error bound $(N\eta)^{-1}$ could be replaced by $(N(\kappa+\eta))^{-1}$. Such an improvement also holds for our multiplicative model. In fact, in this case, we can also improve the convergence rates for the entrywise local laws from $(\sqrt{N\eta})^{-1/2}$ to $N^{-1/2}(\kappa+\eta)^{-1/4}$. Together with these results, we will be able to study the deformed invariant model [11]. These will be studied in our future works. Finally, while we restricted ourselves to the edge local law for the sake of simplicity, the same argument can be used to prove Theorem 2.13 in the bulk, by replacing the spectral domain $\mathcal{D}_\tau(\eta_L,\eta_U)$ with

$$\mathcal{D}_{\text{bulk}} := \{z=E+\mathrm{i}\eta\in\mathbb{C}_+ : E_-+\tau<E<E_+-\tau, \eta_L<\eta<\eta_U\}, \tag{2.27}$$

for any fixed constant $\tau>0$. In fact, the proof will be simpler in this regime. We refer the readers to Remark 3.2 for more details.

Next, we state two important consequences of the local laws: edge eigenvalue rigidity and edge eigenvector delocalization. Denote γ_j as the jth N-quantile (or classical location) of $\mu_\alpha\boxtimes\mu_\beta$ such that

$$\int_{\gamma_j}^{\infty} \mathrm{d}\mu_\alpha\boxtimes\mu_\beta(x)=\frac{j}{N}.$$

Similarly, we denote γ_j^* to be the jth N-quantile of $\mu_A\boxtimes\mu_B$. Recall that $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_N$ are the eigenvalues of $AUBU^*$.

THEOREM 2.15. *Suppose Assumptions 2.2 and 2.4 hold true. For any small constant* $0 < c < 1/2$, *we have that for all* $1 \le i \le cN$,

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

Moreover, the same conclusion holds if γ_i^* *is replaced with* γ_i.

Denote the singular value decomposition (SVD) for Y in (1.1) as

$$Y = \sum_{i=1}^{N} \sqrt{\lambda_i}\, \mathbf{u}_i \mathbf{v}_i^*,$$

where $\{\mathbf{u}\}_i$ and $\{\mathbf{v}_i\}$ are the left and right singular vectors of Y, respectively.

THEOREM 2.16. *Suppose Assumptions 2.2 and 2.4 hold true. For any fixed small constant* $0 < c < 1/2$, *we have that for all* $1 \le i \le cN$,

$$\max_k |\mathbf{u}_i(k)|^2 + \max_\mu |\mathbf{v}_i(\mu)|^2 \prec \frac{1}{N}.$$

REMARK 2.17. We provide some further remarks here. First, in the current paper, the deterministic matrices A and B are both assumed to be positive definite so that $A^{1/2}$ and $B^{1/2}$ are well defined. This ensures the model is symmetric in the sense that A and B can be interchanged freely, and we often use such an argument along our proofs. In particular, we actually use all four matrices in (2.2) and their resolvents; see (4.42) for an illustration, where we apply the Ward identity to the resolvents of $\widetilde{H}$ and $\widetilde{\mathcal{H}}$. Moreover, this symmetry played an important role in [31] and in earlier appearances of subordination functions on which our results relied, for example, [10, 13]. In this sense, even though taking (only) one of A and B to be nonpositive in $H = AUBU^*$ still gives real eigenvalues, some arguments in the present paper cease to work because the model is no longer symmetric and either $\widetilde{H}$ or $\widetilde{\mathcal{H}}$ in (2.2) is not defined. Nonetheless, we believe that the result remains true in this case, and it might be possible to prove with an application of linearization trick. We explain more details on difficulties arising in applications of the linearization trick in the next paragraph.

Second, in [29] the author proved a weak local law for generic self-adjoint polynomials of A and $\widetilde{B}$ (recall $\widetilde{B} = UBU^*$), on the scale of $N^{-1/12}$. This work and its deterministic precursor [12] suggest that an analogous result to the present paper and [7] should hold for generic polynomials. The main idea in [29] was to use the linearization trick to consider the sum of tensors $x_\alpha \otimes A + x_\beta \otimes \widetilde{B}$ for suitably finite and Hermitian matrices x_α, x_β, instead of the given polynomial. Moreover, we point out that this can also apply to the case with a nonpositive matrix, say B, by considering $\sqrt{A}\widetilde{B}\sqrt{A}$ as a polynomial of $\sqrt{A}$ and $\widetilde{B}$. Although it is feasible that the techniques of the current paper and those in [2–4, 7] can apply to the general model in [29], there are two major difficulties in accommodating these arguments to the linearized models. On one hand, it is a nontrivial task to study the limiting distribution and its regularity for the general free polynomial model. In particular, there are no known natural and suitable conditions on μ_α and μ_β like Assumption 2.2, and on the generic polynomials so that the free polynomial is regular, especially near the edge. On the other hand, as will be seen in Section 3.2 below, the proof of local laws relies on many auxiliary scalar quantities; see (3.9) and (3.11) for examples. However, for the general model in [29], due to the linearization, all these quantities should be matrices instead of scalars. Hence, finding the nonscalar equivalents of these auxiliary quantities in the general setting can be challenging. At the current stage we are not aware of a systematic approach to find them in general, except that in the bulk regime some related techniques have been developed for polynomials of Wigner matrices recently in [27].

2.4. *Statistical applications.* In this subsection, we briefly discuss some applications of our results to high-dimensional statistics. First, our results can be used to detect the existence of signals in the signal-plus-noise model when the noise part is of the form $A^{1/2}UBU^*A^{1/2}$. Consider

$$Y = S + Z,$$

where S and Z stand for the signal and noise parts, respectively. Such a model finds important applications in many scientific endeavors. Especially in many cases S is a low-rank symmetric matrix, for example, diffusion tensor imaging (DTI) analysis [42], $\mathbb{Z}_2$ synchronization [30], community detection using stochastic block model (SBM) [1], matrix denoising and recovery [25, 37] and signal processing [43]. While most of the existing literature focuses on the setting that Z is a Wigner matrix, the free multiplicative noise is also considered in the literature [14, 15]. Therefore, we can apply our results to study the signal-plus-noise model when $Z = A^{1/2}UBU^*A^{1/2}$, that is,

$$Y = S + A^{1/2}UBU^*A^{1/2}. \tag{2.28}$$

A fundamental task is to recover the signal matrix S from the observed sample Y in (2.28), and the very first step is to know whether there exists any such signal. From the random matrix theory viewpoint, the eigenvalues of the signal part S can be viewed as outliers which detach from the bulk of $\mu_A \boxtimes \mu_B$. Our Theorem 2.15 can be used to achieve this goal, especially when we can employ the following Onatski's statistic [39]:

$$\mathbb{T} := \max_{1\le i\le C} \frac{\lambda_i(Y) - \lambda_{i+1}(Y)}{\lambda_{i+1}(Y) - \lambda_{i+2}(Y)}, \tag{2.29}$$

where $C > 0$ is some pre-chosen large integer and $\{\lambda_i(Y)\}$ are the eigenvalues of Y which are ordered decreasingly. We now explain how the statistic $\mathbb{T}$ works by considering the simplest case with rank-one alternative. Note that rank-one S already has important applications in $\mathbb{Z}_2$ synchronization [30] and SBM [1]. Formally, we consider the hypothesis test

$$\mathbf{H}_0 : S = 0 \quad \text{versus} \quad \mathbf{H}_a : S = d\mathbf{u}\mathbf{u}^*, \tag{2.30}$$

for some large constant $d > 0$. On one hand, according to Theorem 2.15, under the null hypothesis $\mathbf{H}_0$, the statistic $\mathbb{T}$ should satisfy that $\mathbb{T} = 1 + o_\prec(1)$. On the other hand, if $\mathbf{H}_a$ holds, when d is above some threshold, that is, the signal is relatively strong so that $\lambda_1(Y)$ detaches from the spectrum of $\mu_A \boxtimes \mu_B$, we shall have that with high probability $\mathbb{T} > 1 + \tau$ for some constant $\tau > 0$. Consequently, we can use $\mathbb{T}$ to detect the existence of the signal matrix. We mention that the exact characterization of d, that is, the BBP transition, often requires a sophisticated perturbation argument involving optimal local laws, Theorem 2.13. Moreover, to formally perform the test (2.30), we need to find the distribution of $\mathbb{T}$. Since these are beyond the focuses of the current paper, we defer these problems to future works.

Second, some simple extensions of our results can be applied to provide some insights on the performance of random sketching in the setting of high-dimensional least square regression [24]. Suppose that we observe N data points (x_i, y_i), where $x_i \in \mathbb{R}^p$ are the predictors, and $y_i \in \mathbb{R}$ are the responses. Consider the linear model that $y_i = x_i^\top \beta + \epsilon_i$, where $\beta \in \mathbb{R}^p$ is an unknown parameter and $\epsilon_i' s$ are the white noise error. The ordinary least squares (OLS) estimator for β can be written as

$$\widehat{\beta} = (X^\top X)^{-1} X^\top Y,$$

where $X \in \mathbb{R}^{N\times p}$ collects all x_i and $y \in \mathbb{R}^N$ collects all $y_i, 1 \le i \le N$. The OLS estimator is a gold standard when rank$(X)= p$. However, when both N and p are large, it is computationally expensive to get the OLS estimator. In the high-dimensional setting, sketching is an

effective approach to reduce the size of the problem by multiplying an $M \times N$ matrix S to obtain the sketched data $(\widetilde{X}, \widetilde{Y}) = (SX, SY)$. Then the sketched OLS estimator is [24]

$$\widehat{\beta}_s = (\widetilde{X}^\top \widetilde{X})^{-1} \widetilde{Y}.$$

The computational cost will drop from Np^2 to Mp^2 by introducing the sketching matrix S. The goal of sketching is to find an $M < N$ so that the performance of $\widehat{\beta}$ and $\widehat{\beta}_s$ is similar.

One popular and efficient choice for S is the truncated Haar orthogonal matrix in the sense that S is a submatrix of an $N \times N$ Haar orthogonal matrix. To analyze the performance of $\widehat{\beta}_s$, the core is to study the projection matrix $\widetilde{X}^\top \widetilde{X}$. It is not hard to see that we can rewrite $\mathsf{D}_s := \widetilde{X}^\top \widetilde{X}$ as (see Section A.7 of [24])

$$\mathsf{D}_s = \begin{pmatrix} I_M & 0 \\ 0 & 0 \end{pmatrix} U \begin{pmatrix} \Lambda_p & 0 \\ 0 & 0 \end{pmatrix} U^* \begin{pmatrix} I_M & 0 \\ 0 & 0 \end{pmatrix}, \tag{2.31}$$

where Λ_p is the diagonal matrix containing the nontrivial eigenvalues of $X^\top X$. D_s includes the truncated Haar matrices here due to the block structure of the deterministic matrices.

Based on the above discussion, we see that the core is to analyze the matrix D_s. In fact, the above model is of the form (1.1) by setting

$$A = \begin{pmatrix} I_M & 0 \\ 0 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} \Lambda_p & 0 \\ 0 & 0 \end{pmatrix}. \tag{2.32}$$

For this particular case, even through the positive definite assumptions in Section 2.3 are slightly violated, the results still hold true near the upper edge. These results, especially Theorem 2.13 can be used to establish the convergent rates for the variance efficiency (VE) of $\widehat{\beta}_s$, that is, the increase in parameter estimation error compared to using $\widehat{\beta}$ directly. For instance, together with Theorem 2.3 of [24], we will be able to show that when p is comparable to both M and N

$$\frac{\|\widehat{\beta}_s - \beta\|^2}{\|\widehat{\beta} - \beta\|^2} = \frac{N-p}{M-p} + O_\prec\Big(\frac{1}{\sqrt{N}}\Big). \tag{2.33}$$

(2.33) can be used as a starting point to choose a numerically efficient value of M for the finite sample study. A rigorous discussion is out of the scope of this paper and we will pursue this direction in future works.

Finally, we mention that the key matrices of many other problems are also in the form of (1.2) or $A^{1/2}UBU^*A^{1/2}$. For example, many of the nonlinear kernel-based data acquisition algorithms can be reduced to studying a random matrix of the form (1.2) [19]. Consequently, our results can be potentially applied to check whether common signals have been properly captured by two different sensors using a statistic similar to (2.29) [20]. Moreover, $A^{1/2}UBU^*A^{1/2}$ is also a natural model for spatiotemporal data analysis [15], where A and B are respectively the spatial and temporal covariance matrices. In practice, a spiked model analogous to [22] is more reasonable for real applications. The results in this paper are key ingredients to study such a problem. We will pursue these directions in future works.

3. General structure of the proof.

3.1. *Partial randomness decomposition.* As mentioned earlier, the partial randomness decomposition has been used as an important asset to handle the Haar random matrix in the additive model [4, 5, 7]. For our multiplicative model, we also need to use this tool. The partial randomness decomposition can be regarded as the counterpart of the Schur's complement, which plays a central role in the proof of the local laws [8, 21, 33, 36, 47] when X in (1.1)

has i.i.d. entries. In what follows, we focus on introducing this technique for Haar unitary matrix. We will also briefly discuss how the arguments apply to Haar orthogonal matrix.

Let U be an $N \times N$ Haar unitary random matrix. For all $i \in [\![1, N]\!]$, define $\boldsymbol{v}_i := U\boldsymbol{e}_i$ as the ith column vector of U and θ_i as the argument of $\boldsymbol{e}_i^* \boldsymbol{v}_i$. Following [17], we denote

$$U^{\langle i\rangle} := -\mathrm{e}^{-\mathrm{i}\theta_i} R_i U, \quad \text{where } R_i := I - \boldsymbol{r}_i \boldsymbol{r}_i^*,\ \boldsymbol{r}_i := \sqrt{2}\frac{\boldsymbol{e}_i + \mathrm{e}^{-\mathrm{i}\theta_i}\boldsymbol{v}_i}{\|\boldsymbol{e}_i + \mathrm{e}^{-\mathrm{i}\theta_i}\boldsymbol{v}_i\|}. \tag{3.1}$$

Since $\|\boldsymbol{r}_i\|^2 = 2$, we have that R_i is a Householder reflection. Consequently, $R_i^* = R_i$ and $R_i^2 = I$. Furthermore, it is easy to see that $U^{\langle i\rangle}\boldsymbol{e}_i = \boldsymbol{e}_i$ and $\boldsymbol{e}_i^* U^{\langle i\rangle} = \boldsymbol{e}_i^*$. This implies that $U^{\langle i\rangle}$ is a unitary block-diagonal matrix. In other words, $U_{ii}^{\langle i\rangle} = 1$ and the (i, i)-matrix minor of $U^{\langle i\rangle}$ is Haar distributed on $U(N-1)$ and $\boldsymbol{v}_i$ is uniformly distributed on the $N-1$ unit sphere. Denote

$$\widetilde{B}^{\langle i\rangle} := U^{\langle i\rangle} B \big(U^{\langle i\rangle}\big)^*. \tag{3.2}$$

Since $\boldsymbol{v}_i$ is uniformly distributed on the unit sphere $\mathbb{S}_{\mathbb{C}}^{N-1}$, we can find a Gaussian vector $\widetilde{\boldsymbol{g}}_i \sim \mathcal{N}_{\mathbb{C}}(0, N^{-1} I_N)$ such that

$$\boldsymbol{v}_i = \frac{\widetilde{\boldsymbol{g}}_i}{\|\widetilde{\boldsymbol{g}}_i\|}. \tag{3.3}$$

Armed with the above Gaussian vector, we further define

$$\begin{aligned} &\boldsymbol{g}_i := \mathrm{e}^{-\mathrm{i}\theta_i}\widetilde{\boldsymbol{g}}_i, \qquad \boldsymbol{h}_i := \frac{\boldsymbol{g}_i}{\|\boldsymbol{g}_i\|} = \mathrm{e}^{-\mathrm{i}\theta_i}\boldsymbol{v}_i, \qquad \ell_i := \frac{\sqrt{2}}{\|\boldsymbol{e}_i + \boldsymbol{h}_i\|}, \\ &\mathring{\boldsymbol{g}}_i := \boldsymbol{g}_i - g_{ii}\boldsymbol{e}_i, \qquad \mathring{\boldsymbol{h}}_i := \boldsymbol{h}_i - h_{ii}\boldsymbol{e}_i. \end{aligned} \tag{3.4}$$

It is easy to see from (3.1) that

$$\boldsymbol{r}_i = \ell_i(\boldsymbol{e}_i + \boldsymbol{h}_i), \qquad R_i\boldsymbol{e}_i = -\boldsymbol{h}_i, \qquad R_i\boldsymbol{h}_i = -\boldsymbol{e}_i, \tag{3.5}$$

which further implies that

$$\boldsymbol{h}_i^* \widetilde{B}^{\langle i\rangle} R_i = -\boldsymbol{e}_i^* \widetilde{B}, \qquad \boldsymbol{e}_i^* \widetilde{B}^{\langle i\rangle} R_i = -\boldsymbol{h}_i^* \widetilde{B} = -b_i \boldsymbol{h}_i^*. \tag{3.6}$$

The above equations provide several convenient identities for the Haar unitary matrix. Moreover, since $\widetilde{B}^{\langle i\rangle}$ is independent of both $\boldsymbol{h}_i$ and R_i, we can establish accurate large deviation estimates for quantities related to (3.6); see Section I.2 of [18] for more details. Finally, for the Haar orthogonal random matrix on $O(N)$, the only difference lies in the partial randomness decomposition. In fact, we can decompose an orthogonal matrix U in the same way as in (3.1), except that the factor $\mathrm{e}^{-\mathrm{i}\theta_i}$ in (3.1) should be replaced by $\mathrm{sgn}(\boldsymbol{e}_i^*\boldsymbol{v}_i)$. We refer the readers to [3], Appendix A, for more details.

3.2. *Sketch of the proof route.* In this subsection, we summarize the main route of the proof. Our proof basically follows [7] and we focus on explaining how to adapt their proof strategies to study the multiplicative model. For a proof route of the additive model, we refer the readers to [7], Section 4.2.

Recall (2.1). Without loss of generality, till the end of the paper, we assume that both A and B are normalized such that $\mathrm{tr}\, A = \mathrm{tr}\, B = 1$. First, we introduce the random equivalents of the subordination functions Ω_A and Ω_B in terms of the resolvents. They are the starting points of the arguments and the counterparts of the additive model as in equation (5.2) of [7].

DEFINITION 3.1 (Approximate subordination functions). For $z \in \mathbb{C} \setminus \mathbb{R}_+$, we define

$$\Omega_A^c \equiv \Omega_A^c(z) := \frac{z \operatorname{tr} \widetilde{B} G}{1 + z \operatorname{tr} G}, \qquad \Omega_B^c \equiv \Omega_B^c(z) := \frac{z \operatorname{tr} A G}{1 + z \operatorname{tr} G} = \frac{z \operatorname{tr} \mathcal{G} \widetilde{A}}{1 + z \operatorname{tr} \mathcal{G}}.$$

As mentioned earlier in Section 1, while the final goal is to prove (2.22), following the strategy of [7], we actually work with the approximate subordination functions in Definition 3.1 and several auxiliary quantities. To identify these auxiliary quantities, we need several crucial decompositions. By replacing Ω_B with Ω_B^c in (2.22), we observe that

$$\begin{aligned} &(zG_{ii} + 1) - \frac{a_i}{a_i - \Omega_B^c} \\ &\quad = a_i (\widetilde{B} G)_{ii} - \frac{a_i}{a_i - \Omega_B^c} \\ &\quad = \frac{a_i}{(1 + z \operatorname{tr} G)(a_i - \Omega_B^c)} \big((z \operatorname{tr} G + 1)(a_i - \Omega_B^c)(\widetilde{B} G)_{ii} - (z \operatorname{tr} G + 1) \big) \\ &\quad = \frac{a_i z}{(1 + z \operatorname{tr} G)(a_i - \Omega_B^c)} \big(G_{ii} \operatorname{tr}(A \widetilde{B} G) - \operatorname{tr}(GA)(\widetilde{B} G)_{ii} \big), \end{aligned} \tag{3.7}$$

where we used the elementary identity $(HG)_{ii} - zG_{ii} = a_i (\widetilde{B} G)_{ii} - zG_{ii} = 1$. In light of Proposition 2.11 and (3.7), to prove (2.22), it suffices to control

$$Q_i := G_{ii} \operatorname{tr}(A \widetilde{B} G) - \operatorname{tr}(GA)(\widetilde{B} G)_{ii}, \tag{3.8}$$

and show that Ω_B and Ω_B^c are sufficiently close. Note that Q_i is the counterpart of equation (4.11) of [7].

We first present the detailed decomposition of Q_i. In particular, following [7], Section 4.2, we discuss how to decompose and explore the independence structure of $(\widetilde{B} G)_{ii}$ using the partial randomness decomposition. Using (3.2) and (3.4), we introduce the notation

$$S_i := \boldsymbol{h}_i^* \widetilde{B}^{\langle i \rangle} G \boldsymbol{e}_i, \qquad \mathring{S}_i := \mathring{\boldsymbol{h}}_i^* \widetilde{B}^{\langle i \rangle} G \boldsymbol{e}_i, \qquad T_i := \boldsymbol{h}_i^* G \boldsymbol{e}_i = \mathrm{e}^{\mathrm{i}\theta_i} \boldsymbol{e}_i^* U^* G \boldsymbol{e}_i, \ \mathring{T}_i := \mathring{\boldsymbol{h}}_i^* G \boldsymbol{e}_i, \tag{3.9}$$

where we used $(e^{-\mathrm{i}\theta_i})^* = e^{\mathrm{i}\theta_i}$. By the construction of $U^{\langle i \rangle}$ in (3.1) and (3.5), we find that

$$(\widetilde{B} G)_{ii} = -\boldsymbol{h}_i^* \widetilde{B}^{\langle i \rangle} R_i G \boldsymbol{e}_i.$$

Using the definition of R_i in (3.1) and $\boldsymbol{r}_i = \ell_i (\boldsymbol{e}_i + \boldsymbol{h}_i)$, we have the following expansion:

$$(\widetilde{B} G)_{ii} = -\boldsymbol{h}_i^* \widetilde{B}^{\langle i \rangle} G \boldsymbol{e}_i + \ell_i^2 \boldsymbol{h}_i^* \widetilde{B}^{\langle i \rangle} (\boldsymbol{e}_i + \boldsymbol{h}_i)(\boldsymbol{e}_i + \boldsymbol{h}_i)^* G \boldsymbol{e}_i.$$

Moreover, utilizing the notation in (3.9), we can write that

$$(\widetilde{B} G)_{ii} = -S_i + \ell_i^2 (\boldsymbol{h}_i^* \widetilde{B}^{\langle i \rangle} \boldsymbol{e}_i + \boldsymbol{h}_i^* \widetilde{B}^{\langle i \rangle} \boldsymbol{h}_i)(G_{ii} + T_i).$$

Further, since R_i is a projection satisfying (3.5), we have that

$$\begin{aligned} (\widetilde{B} G)_{ii} &= -S_i + \ell_i^2 (-b_i \boldsymbol{h}_i^* R_i \boldsymbol{h}_i + \boldsymbol{h}_i^* \widetilde{B}^{\langle i \rangle} \boldsymbol{h}_i)(G_{ii} + T_i) \\ &= -S_i + \ell_i^2 (b_i h_{ii} + \boldsymbol{h}_i^* \widetilde{B}^{\langle i \rangle} \boldsymbol{h}_i)(G_{ii} + T_i). \end{aligned} \tag{3.10}$$

We will see later in our proof (e.g., (4.23)), the discussion boils down to controlling S_i and T_i, which are the counterparts of equation (4.10) of [7].

In the actual proof, inspired by the arguments in [7], instead of working directly with S_i and T_i, we deal with the following quantities:

$$\begin{aligned} P_i &:= Q_i + (G_{ii} + T_i) \Upsilon, \\ K_i &:= T_i + \operatorname{tr}(GA) \big(b_i T_i + (\widetilde{B} G)_{ii} \big) - \operatorname{tr}(GA \widetilde{B})(G_{ii} + T_i), \end{aligned} \tag{3.11}$$

where Υ is defined as

$$\Upsilon := \big(\operatorname{tr}(GA\widetilde{B})\big)^2 - \operatorname{tr}(GA)\operatorname{tr}(\widetilde{B}GA\widetilde{B}) - \operatorname{tr}(GA\widetilde{B}) + \operatorname{tr}(GA). \tag{3.12}$$

The analogs of the above quantities for the additive model have been defined in equations (4.12), (4.14) and (4.15) of [7]. On one hand, P_i and K_i are closely related to S_i and T_i. On the other hand, they are easily handled. In fact, using $GA\widetilde{B} = A\widetilde{B}G = zG + I$ and $\operatorname{tr} B = \operatorname{tr}\widetilde{B} = 1$, we recognize that Υ is an average of Q_i, that is,

$$\begin{aligned}\Upsilon &= \operatorname{tr}(A\widetilde{B}G)\big(\operatorname{tr}(zG+I)-1\big) - \operatorname{tr}(GA)\big(\operatorname{tr}\big(\widetilde{B}(zG+I)\big) - \operatorname{tr}\widetilde{B}\big)\\ &= z\big(\operatorname{tr}(G)\operatorname{tr}(A\widetilde{B}G) - \operatorname{tr}(GA)\operatorname{tr}(\widetilde{B}G)\big) = \frac{z}{N}\sum_i Q_i.\end{aligned} \tag{3.13}$$

The proof of the estimates of the above quantities relies on a two-level approach, which is commonly used in the proofs of local laws for random matrices. On the first level, we provide bounds for a fixed spectral parameter z under the condition that G_{ii}, $\mathcal{G}_{ii}$ and T_i satisfy a weak a priori bound (cf. Assumption 4.1). On the second level, we will verify the above a priori bound and further prove they hold uniformly in z.

On the first level, the proof strategy contains three steps. In Step 1, we establish recursive estimates for the high moments of P_i and K_i, that is, Proposition 4.2, which is the counterpart of Proposition 5.1 of [7]. The main idea is to employ the (Gaussian) integration by parts with respect to the coordinates of $\boldsymbol{h}_i$ since $\widetilde{B}^{\langle i\rangle}$ is independent of $\boldsymbol{h}_i$. This step concludes that all the quantities P_i, K_i, T_i, Q_i and Υ can be bounded by $(N\eta)^{-1/2}$. The actual proofs will be presented in Section 4. In Step 2, we derive a rough bound on the averaged quantities. Especially, since Q_i is the most fundamental quantity, we focus on the form $N^{-1}\sum d_i Q_i$, where d_i's are some generic weights. The proof also concerns the recursive moment estimates as in Step 1. This step yields that the averaged quantity can be bounded by $\sqrt{\operatorname{Im} m_H(z)}(N\eta)^{-1}$, which improves the bounds from Step 1. The arguments will be given in Section B.1 of [18]. In Step 3, we prove that for some specific weights d_i (cf. (B.31) and (B.32)), the averaged quantities can be bounded by $\operatorname{Im} m_H(z)(N\eta)^{-1}$. Note that the weights we choose here are the counterparts of equation (7.12) of [7]. As a byproduct, we obtain a priori the bound for $|\Omega_B - \Omega_B^c|$ and $|\Omega_A - \Omega_A^c|$. In fact, controlling the differences above can also be reduced to Υ due to the following decomposition:

$$\begin{aligned}\Omega_A^c\Omega_B^c - zM_{\mu_H}(z) &= \frac{z^2}{(1+zm_H(z))^2}\big(\operatorname{tr}(GA)\operatorname{tr}(\widetilde{B}G) - \operatorname{tr}(G)\operatorname{tr}(A\widetilde{B}G)\big)\\ &= -\frac{z}{(1+zm_H(z))^2}\Upsilon(z),\end{aligned} \tag{3.14}$$

where we used $A\widetilde{B}G = zG + I$. All these will be discussed in Section B.2 of [18].

On the second level, the proof consists of two parts. In Part 1, we establish the *weak local laws* by verifying Assumption 4.1 and prove the uniformity of the estimates in z is obtained by a continuity argument. The weak local laws state that most of the aforementioned quantities, for example, P_i, K_i and (3.7) can be bounded by $(N\eta)^{-1/2}$ uniformly in z. Moreover, $|\Omega_B - \Omega_B^c|$ and $|\Omega_A - \Omega_A^c|$ can be uniformly bounded by $(N\eta)^{-1/3}$. The formal arguments are given in Section C.1 of [18]. In Part 2, using the weak local laws, we complete the proof of Theorem 2.13, which is referred to as *strong local laws*. The arguments can be found in Section C.2 of [18]. Finally, we emphasize that even though the above strategy is sketched for the diagonal entries, the off-diagonal entries can be handled similarly. We discuss this aspect in detail at the end of Section C.1 of [18].

Once Theorem 2.13 is proved, other theorems can be justified based on it. First, Theorems 2.15 can be proved by translating the closeness of the resolvent into the closeness of the eigenvalues and the quantiles of $\mu_A \boxtimes \mu_B$ using the Helffer–Sjöstrand formula, and Theorem 2.16 can be proved by exploring the imaginary part of the resolvents.

REMARK 3.2. Before concluding this section, we briefly discuss how the above strategy can be used to obtain the bulk local laws when the spectral parameter z is in $\mathcal{D}_{\text{bulk}}$ defined in (2.27). Similar results have been proved for the additive model in [4].

In fact, in the bulk regime, the proof will be easier. The main reason is that when $z \in \mathcal{D}_{\text{bulk}}$, the key parameter $\kappa \equiv \kappa(z) := \min\{|z - E_-|, |z - E_+|\}$ satisfies $\kappa \sim 1$ so that $\sqrt{\kappa + \eta} \sim 1$. Consequently, according to [31], in contrast to Proposition A.6 of our supplement [18], the key quantities can be controlled more easily in the sense that $\operatorname{Im} m_{\mu_\alpha \boxtimes \mu_\beta}(z) \sim 1$ and $|\mathcal{S}_{\alpha\beta}(z)| \sim 1$. Combining these updates with lines of the proof of Proposition 2.11, we can update the results of Proposition 2.11 by inserting $\kappa \sim 1$.

Next, we explain how the two-level approach applies and is easier. On the first level, the bulk regime only needs Steps 1 and 2. The reason is that since when $z \in \mathcal{D}_{\text{bulk}}$ we have $\operatorname{Im} m_{\mu_A \boxtimes \mu_B}(z) \sim 1$, Steps 1 and 2 will prove that $\operatorname{Im} m_H(z) \sim 1$ and the averaged quantities will therefore be bounded by $(N\eta)^{-1}$ which is already optimal. On the second level, in contrast to the edge regime where we have to decompose the edge spectral domain according to different scales of $\sqrt{\kappa + \eta}$ as in Section C.2 of our supplement [18], we can work on the whole bulk spectral domain directly as $\kappa \sim 1$.

4. Entrywise resolvent subordination. In this section, we prove a subordination property for the resolvent entries, that is, Proposition 4.2. In particular, we prove (2.22) and (2.25) of Theorem 2.13 and other related quantities for fixed spectral parameter z with a priori bound, that is, Assumption 4.1. This completes Step 1 of the first level of our proof route as summarized in Section 3.2. Moreover, the proof of Proposition 4.2, especially the recursive moment estimates in Lemma 4.3, is a representative formal argument of our proof strategies in the sense that Steps 2 and 3 follow a similar discussion. Analogous arguments have been made for the additive model in Section 5 of [7].

We first introduce the assumptions. Denote

$$\Lambda_{di} := \left| zG_{ii} + 1 - \frac{a_i}{a_i - \Omega_B} \right|, \qquad \Lambda_d := \max_i \Lambda_{di}, \qquad \Lambda_T := \max_i |T_i|. \tag{4.1}$$

Similarly, we define Λ_{di}^c and Λ_d^c by replacing Ω_B with its approximate Ω_B^c. Moreover, denote $\widetilde{\Lambda}_{di}$, $\widetilde{\Lambda}_d$ and $\widetilde{\Lambda}_T$ as

$$\widetilde{\Lambda}_{di} := \left| z\mathcal{G}_{ii} + 1 - \frac{b_i}{b_i - \Omega_A} \right|, \qquad \widetilde{\Lambda}_d := \max_i \widetilde{\Lambda}_{di}, \qquad \widetilde{\Lambda}_T := \max_i |\boldsymbol{e}_i^* U \mathcal{G} \boldsymbol{e}_i|.$$

Furthermore, $\widetilde{\Lambda}_{di}^c$ and $\widetilde{\Lambda}_d^c$ are defined by replacing Ω_A with Ω_A^c. In this section, the statements and proofs are based on Assumption 4.1, which provides a priori the bound for the essential quantities. It will be verified in Section C.1 of [18].

ASSUMPTION 4.1. Recall (2.20). For the small constant $\gamma > 0$ in (2.21), fix $z \in \mathcal{D}_\tau(\eta_L, \eta_U)$, we suppose the following hold true:

$$\Lambda_d(z) \prec N^{-\gamma/4}, \qquad \widetilde{\Lambda}_d(z) \prec N^{-\gamma/4}, \qquad \Lambda_T \prec 1, \qquad \widetilde{\Lambda}_T \prec 1. \tag{4.2}$$

We now state the main result of this section, Proposition 4.2, that provides the estimates for the diagonal entries of the resolvents and is an analogue of Proposition 5.1 of [7] for the additive model. The arguments of the off-diagonal entries are similar and we refer the readers to the discussion at the end of Section C.1 of [18] for more details. Throughout the paper, we will consistently use the following control parameter:

$$\Psi \equiv \Psi(z) := \sqrt{\frac{1}{N\eta}}, \qquad \Pi_i \equiv \Pi_i(z) := \sqrt{\frac{\operatorname{Im} G_{ii}(z) + \operatorname{Im} \mathcal{G}_{ii}(z)}{N\eta}}. \tag{4.3}$$

PROPOSITION 4.2. *Recall* (3.11). *Suppose that the assumptions of Theorem* 2.13 *and Assumption* 4.1 *hold. Fix* $z \in \mathcal{D}_\tau(\eta_L, \eta_U)$. *For all* $i \in [\![1, N]\!]$, *we have that*

$$|P_i(z)| \prec \Psi(z), \qquad |K_i(z)| \prec \Psi(z). \tag{4.4}$$

Furthermore, we have

$$\Lambda_d^c(z) \prec \Psi(z), \quad \Lambda_T \prec \Psi(z), \qquad \widetilde{\Lambda}_d^c \prec \Psi(z), \qquad \widetilde{\Lambda}_T \prec \Psi(z), \qquad \Upsilon \prec \Psi(z). \tag{4.5}$$

In the rest of this section, we follow the proof strategy of [7], Proposition 5.1, to prove Proposition 4.2. The following resolvent identities will be frequently used in the proof:

$$\begin{aligned} (HG)_{ii} - zG_{ii} &= a_i(\widetilde{B}G)_{ii} - zG_{ii} = 1, \\ (\mathcal{G}\mathcal{H})_{ii} - z\mathcal{G}_{ii} &= b_i(\mathcal{G}\widetilde{A})_{ii} - z\mathcal{G}_{ii} = 1. \end{aligned} \tag{4.6}$$

PROOF OF PROPOSITION 4.2. We first prove (4.5) assuming (4.4) holds. The arguments rely on the estimate

$$T_i \prec N^{-\gamma/4}, \tag{4.7}$$

where $\gamma > 0$ is introduced in (4.2).

Before proving (4.5), we pause to justify (4.7). By (4.4) and (3.11), we have

$$T_i\big(1 + b_i \operatorname{tr}(GA) - \operatorname{tr}(GA\widetilde{B})\big) = \operatorname{tr}(GA\widetilde{B})G_{ii} - \operatorname{tr}(GA)(\widetilde{B}G)_{ii} + \mathrm{O}_\prec(\Psi). \tag{4.8}$$

By (4.6), we have that

$$(\widetilde{B}G)_{ii} = \frac{zG_{ii} + 1}{a_i}, \qquad G_{ii} = \frac{1}{z}\big(a_i(\widetilde{B}G)_{ii} - 1\big). \tag{4.9}$$

Based on (4.9), on one hand, by (4.2), we find that

$$(\widetilde{B}G)_{ii} = \frac{1}{a_i - \Omega_B} + \mathrm{O}_\prec\big(N^{-\gamma/4}\big). \tag{4.10}$$

On the other hand, using the fact that $\operatorname{tr}(GA\widetilde{B}) = zm_H(z) + 1$ and a relation similar to (2.10), together with (4.2), we conclude that

$$\operatorname{tr}(GA\widetilde{B}) = \int \frac{x}{x - \Omega_B}\,\mathrm{d}\mu_A(x) + \mathrm{O}_\prec\big(N^{-\gamma/4}\big) = zm_{\mu_A \boxtimes \mu_B}(z) + 1 + \mathrm{O}_\prec\big(N^{-\gamma/4}\big), \tag{4.11}$$

and

$$\begin{aligned} \operatorname{tr}(GA) &= \frac{1}{N}\sum_i a_i G_{ii} = \frac{\Omega_B}{z}\frac{1}{N}\sum_i \frac{a_i}{a_i - \Omega_B} + \mathrm{O}_\prec\big(N^{-\gamma/4}\big) \\ &= \frac{\Omega_B}{z}\big(zm_{\mu_A \boxtimes \mu_B}(z) + 1\big) + \mathrm{O}_\prec\big(N^{-\gamma/4}\big). \end{aligned} \tag{4.12}$$

Combining (4.8), (4.10), (4.11) and (4.12), using (4.2), we have that

$$T_i\left(1 + \big(zm_{\mu_A \boxtimes \mu_B}(z) + 1\big)\left(\frac{b_i\Omega_B}{z} - 1\right)\right) = \mathrm{O}_\prec\big(\Psi + N^{-\gamma/4}\big). \tag{4.13}$$

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

$$\begin{aligned} 1 + \big(zm_{\mu_A \boxtimes \mu_B}(z) + 1\big)\left(\frac{b_i\Omega_B}{z} - 1\right) &= \big(zm_{\mu_A \boxtimes \mu_B}(z) + 1\big)\left(\frac{b_i\Omega_B}{z} - M_{\mu_A \boxtimes \mu_B}(z)\right) \\ &= \big(zm_{\mu_A \boxtimes \mu_B}(z) + 1\big)\frac{\Omega_B}{z}(b_i - \Omega_A). \end{aligned} \tag{4.14}$$

By (4.14), a relation similar to (2.10), and (i) of Proposition 2.11, we have proved the claim (4.7) using (4.13).

Then we prove (4.5). First, using (4.4) and the definitions in (3.11), we find that

$$\frac{1}{N}\sum_i a_i P_i = \Upsilon \frac{1}{N}\sum_i a_i (G_{ii} + T_i) \prec \Psi, \tag{4.15}$$

where we used the fact that $\{a_i\}$ are bounded. By (4.15), (4.12) and (i) of Proposition 2.11, we have proved that

$$\Upsilon \prec \Psi. \tag{4.16}$$

Second, using the definition of P_i in (3.11), the expansion (3.7), and (4.16), we have proved that $\Lambda_d^c \prec \Psi$. Third, by (4.8) and a discussion similar to (4.7) with the bound $\Lambda_d^c \prec \Psi$, it is easy to see that $\Lambda_T \prec \Psi(z)$. Finally, the proof for $\widetilde{\Lambda}_d^c$ and $\widetilde{\Lambda}_T$ follows from an argument similar to (3.7) and (4.6). This completes the proof of (4.5).

It remains to prove (4.4), which is equivalent to the bounds for high moments of P_i and K_i. More specifically, by the Markov inequality, it suffices to prove for all positive integer $p \geq 2$, that the following hold:

$$\mathbb{E}\big[|P_i|^{2p}\big] \prec \Psi^{2p} \quad \text{and} \quad \mathbb{E}\big[|K_i|^{2p}\big] \prec \Psi^{2p}. \tag{4.17}$$

The proof of (4.17) makes use of the recursive estimates, that is, Lemma 4.3 below. Denote

$$\mathfrak{X}_i^{(p,q)} := P_i^p \overline{P_i^q} \quad \text{and} \quad \mathfrak{Y}_i^{(p,q)} := K_i^p \overline{K_i^q}. \tag{4.18}$$

LEMMA 4.3. *For any fixed integer $p \geq 2$ and $i \in [\![1, N]\!]$, we have that*

$$\mathbb{E}\big[\mathfrak{X}_i^{(p,p)}\big] \leq \mathbb{E}\big[\mathrm{O}_\prec(\Psi)\mathfrak{X}_i^{(p-1,p)} + \mathrm{O}_\prec(\Psi^2)\mathfrak{X}_i^{(p-2,p)} + \mathrm{O}_\prec(\Psi^2)\mathfrak{X}_i^{(p-1,p-1)}\big], \tag{4.19}$$

$$\mathbb{E}\big[\mathfrak{Y}_i^{(p,p)}\big] \leq \mathbb{E}\big[\mathrm{O}_\prec(\Psi)\mathfrak{Y}_i^{(p-1,p)} + \mathrm{O}_\prec(\Psi^2)\mathfrak{Y}_i^{(p-2,p)} + \mathrm{O}_\prec(\Psi^2)\mathfrak{Y}_i^{(p-1,p-1)}\big]. \tag{4.20}$$

We next explain how Lemma 4.3 implies (4.17) and will give the proof of Lemma 4.3 at the end of this section. Recall that for any positive numbers $\mathsf{u}, \mathsf{v} > 0$ we have

$$\mathsf{u}\mathsf{v} \leq \frac{\mathsf{u}^m}{m} + \frac{\mathsf{v}^n}{n} \quad \text{where } m, n > 1 \text{ are real numbers with } \frac{1}{m} + \frac{1}{n} = 1. \tag{4.21}$$

For $k = 1, 2$, any arbitrary small constant $\epsilon > 0$ and any random variable $\mathfrak{N} = \mathrm{O}_\prec(\Psi^k)$ satisfying $\mathbb{E}[|\mathfrak{N}|^q] \prec \Psi^{qk}$, we have that

$$\begin{aligned}
\mathbb{E}\big[|\mathfrak{N} P_i^{2p-k}|\big] &= \mathbb{E}\big[|N^\epsilon \mathfrak{N}||N^{-\frac{\epsilon}{2p-k}} P_i|^{2p-k}\big] \\
&\leq \frac{k N^{\frac{2p\epsilon}{k}}}{2p}\mathbb{E}\big[|\mathfrak{N}|^{\frac{2p}{k}}\big] + \frac{(2p-k)N^{-\frac{2p\epsilon}{(2p-k)^2}}}{2p}\mathbb{E}\big[|P_i|^{(2p-k)\frac{2p}{2p-k}}\big] \\
&\leq \frac{k N^{(\frac{2p}{k}+1)\epsilon}}{2p}\Psi^{2p} + \frac{(2p-k)N^{-\frac{2p\epsilon}{(2p-k)^2}}}{2p}\mathbb{E}\big[|P_i|^{2p}\big],
\end{aligned} \tag{4.22}$$

where in the first inequality we used (4.21) with $m = 2p/k$ and $n = 2p/(2p-k)$, and in the second inequality we used $\mathbb{E}[|\mathfrak{N}|^q] \prec \Psi^{qk}$. Together with (4.19), it yields that

$$\mathbb{E}\big[|P_i|^{2p}\big] \leq \frac{3}{2p} N^{(2p+1)\epsilon}\Psi^{2p} + \frac{3(2p-1)}{2p} N^{-\frac{2p\epsilon}{(2p-1)^2}}\mathbb{E}\big[|P_i|^{2p}\big].$$

Since $\epsilon > 0$ is arbitrarily small, we can conclude the first part of (4.17). The second part can be proved similarly and we omit the details here. This completes the proof of (4.4) and hence the proof of Proposition 4.2. □

The rest of this subsection is devoted to the proof of Lemma 4.3. This type of estimates have been used in Lemma 5.2 of [7] to study the analogous quantities for the additive model and our arguments basically follow the proof therein. In particular, similar to equation (5.34) of [7], integration by parts will be mainly applied to the term $\mathring{S}_i$ in (4.30) to generate several hidden terms which will cancel many other existing larger order terms. Throughout the proof, we will need some derivative formulas and large deviation estimates as our technical inputs. These can be found in Lemmas I.1–I.4 of our supplement [18].

PROOF OF LEMMA 4.3. We start with the proof of (4.19). Recall (3.9). Since $h_{ii}\boldsymbol{e}_i^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i = b_i h_{ii}G_{ii}$, we can rewrite (3.10) as

$$(\widetilde{B}G)_{ii} = -S_i + \ell_i^2\big(b_i h_{ii} + \boldsymbol{h}_i^*\widetilde{B}^{\langle i\rangle}\boldsymbol{h}_i\big)(G_{ii}+T_i) = -\mathring{S}_i + G_{ii} + T_i + \mathsf{e}_{i1}, \tag{4.23}$$

where we denoted

$$\mathsf{e}_{i1} := (\ell_i^2 - 1)b_i h_{ii}G_{ii} + \big(\ell_i^2\boldsymbol{h}_i^*\widetilde{B}^{\langle i\rangle}\boldsymbol{h}_i - 1\big)(G_{ii}+T_i) + \ell_i^2 b_i h_{ii}T_i. \tag{4.24}$$

Recall $\widetilde{\boldsymbol{g}} \sim \mathcal{N}_{\mathbb{C}}(0, N^{-1}I_N)$. By Lemma I.2 of [18], we see that

$$h_{ii} = \|\widetilde{\boldsymbol{g}}_i\|^{-1}|\boldsymbol{e}_i^*\widetilde{\boldsymbol{g}}_i| \prec N^{-1/2}. \tag{4.25}$$

Consequently, using the definitions in (3.4), we obtain that

$$\ell_i^2 = \frac{2}{\|\boldsymbol{e}_i + \boldsymbol{h}_i\|^2} = \frac{1}{1+\boldsymbol{e}_i^*\boldsymbol{h}_i} = 1 + \mathrm{O}_{\prec}(N^{-1/2}). \tag{4.26}$$

Moreover, by (3.1), (3.5), (3.3) and Lemma I.2 of [18], we have

$$\boldsymbol{h}_i^*\widetilde{B}^{\langle i\rangle}\boldsymbol{h}_i = \boldsymbol{h}_i^* R_i\widetilde{B}R_i\boldsymbol{h}_i = \boldsymbol{e}_i^*\widetilde{B}\boldsymbol{e}_i = \frac{1}{\|\widetilde{\boldsymbol{g}}_i\|^2}\widetilde{\boldsymbol{g}}_i^* B\widetilde{\boldsymbol{g}}_i = 1 + \mathrm{O}_{\prec}(N^{-1/2}), \tag{4.27}$$

where we recall that B is normalized such that $\operatorname{tr} B = 1$. Using the definition (4.24), by (4.25), (4.26), (4.27) and (4.2), we conclude that

$$|\mathsf{e}_{i1}| \prec N^{-1/2}. \tag{4.28}$$

Therefore, by (3.11), (4.23) and (4.28), we have shown that

$$\begin{aligned}\mathbb{E}\big[\mathfrak{X}_i^{(p,p)}\big] = \mathbb{E}\big[&\big(G_{ii}\operatorname{tr}(A\widetilde{B}G) + \operatorname{tr}(GA)(\mathring{S}_i) + (G_{ii}+T_i)\big(\Upsilon - \operatorname{tr}(GA)\big)\big)\mathfrak{X}_i^{(p-1,p)}\big] \\ &+ \mathbb{E}\big[\mathsf{e}_{i1}\operatorname{tr}(GA)\mathfrak{X}_i^{(p-1,p)}\big].\end{aligned} \tag{4.29}$$

Next, we control all the terms on the RHS of (4.29). We mainly focus on the term involving $\operatorname{tr}(GA)(\mathring{S}_i)$. As we will see later, by exploring the hidden terms using integration by parts, the term involving $\operatorname{tr}(GA)(\mathring{S}_i)$ will generate several terms which would cancel the rest of the terms on the RHS of (4.29) algebraically. Note that

$$\mathring{S}_i = \mathring{\boldsymbol{h}}_i^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i = \sum_k \mathring{\boldsymbol{h}}_i^*\boldsymbol{e}_k\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i = \sum_k^{(i)}\overline{g}_{ik}\frac{1}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i, \tag{4.30}$$

where we use the shorthand notation $\sum_k^{(i)}$ to represent the sum over $[\![1,N]\!]\backslash\{i\}$. Our calculation relies on the following integration by parts formula for $g \sim \mathcal{N}(0,\sigma^2)$ (see equation (5.33) of [7]):

$$\int_{\mathbb{C}} \bar{g}f(g,\bar{g})e^{-\frac{|g|^2}{\sigma^2}}\,\mathrm{d}^2 g = \sigma^2\int_{\mathbb{C}}\partial_g f(g,\bar{g})e^{-\frac{|g|^2}{\sigma^2}}\,\mathrm{d}^2 g, \tag{4.31}$$

where $f:\mathbb{C}^2\to\mathbb{C}$ is a differentiable function. By (4.30) and (4.31), we have

$$
\begin{aligned}
\mathbb{E}\big[\mathring{S}_i \operatorname{tr}(GA)\mathfrak{X}^{(p-1,p)}\big] &= \sum_k^{(i)} \mathbb{E}\bigg[\overline{g}_{ik}\frac{1}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i \operatorname{tr}(GA)\mathfrak{X}^{(p-1,p)}\bigg] \\
&= \frac{1}{N}\sum_k^{(i)} \mathbb{E}\bigg[\frac{\partial\|\boldsymbol{g}_i\|^{-1}}{\partial g_{ik}}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i \operatorname{tr}(GA)\mathfrak{X}^{(p-1,p)}\bigg] \\
&\quad + \frac{1}{N}\sum_k^{(i)} \mathbb{E}\bigg[\frac{1}{\|\boldsymbol{g}_i\|}\frac{\partial(\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i)}{\partial g_{ik}} \operatorname{tr}(GA)\mathfrak{X}^{(p-1,p)}\bigg] \\
&\quad + \frac{1}{N}\sum_k^{(i)} \mathbb{E}\bigg[\frac{\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i}{\|\boldsymbol{g}_i\|}\frac{\partial \operatorname{tr}(GA)}{\partial g_{ik}}\mathfrak{X}^{(p-1,p)}\bigg] \\
&\quad + \frac{p-1}{N}\mathbb{E}\bigg[\frac{1}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i \operatorname{tr}(GA)\frac{\partial P_i}{\partial g_{ik}}\mathfrak{X}^{(p-2,p)}\bigg] \\
&\quad + \frac{p}{N}\mathbb{E}\bigg[\frac{1}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i \operatorname{tr}(GA)\frac{\partial\overline{P}_i}{g_{ik}}\mathfrak{X}^{(p-1,p-1)}\bigg],
\end{aligned}
\tag{4.32}
$$

where we recall that $\mathfrak{X}^{(p-1,p)} = P_i^{p-1}\overline{P}_i^{p}$ as defined in (4.18). We point out that the second term on the RHS of (4.32) will provide some hidden terms for cancellation. Further, since $\widetilde{B}^{\langle i\rangle}$ is independent of $\boldsymbol{v}_i$, we have that

$$
\frac{\partial(\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i)}{\partial g_{ik}} = \boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}\frac{\partial G}{\partial g_{ik}}\boldsymbol{e}_i.
$$

We start with the analysis of the household reflection defined in (3.1). Recall $\boldsymbol{r}_i = \ell_i(\boldsymbol{e}_i + \boldsymbol{h}_i)$ and ℓ_i defined in (3.4). By (I.1), (I.2) and (I.4) of [18], we have that

$$
\begin{aligned}
\frac{\partial R_i}{\partial g_{ik}} &= -\ell_i^4\|\boldsymbol{g}_i\|^{-1}\overline{h}_{ik}h_{ii}(\boldsymbol{e}_i+\boldsymbol{h}_i)(\boldsymbol{e}_i+\boldsymbol{h}_i)^* \\
&\quad - \ell_i^2\|\boldsymbol{g}_i\|^{-1}\big(\boldsymbol{e}_k\boldsymbol{e}_i^* - \overline{h}_{ik}\big(\boldsymbol{h}_i\boldsymbol{e}_i^* + \boldsymbol{e}_i\boldsymbol{h}_i^*\big) + \boldsymbol{e}_k\boldsymbol{h}_i^* - \overline{h}_{ik}\boldsymbol{h}_i\boldsymbol{h}_i^* - \overline{h}_{ik}\boldsymbol{h}_i\boldsymbol{h}_i^*\big).
\end{aligned}
$$

We can further rewrite the above equation as

$$
\frac{\partial R_i}{\partial g_{ik}} = -\frac{\ell_i^2}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k\big(\boldsymbol{e}_i^* + \boldsymbol{h}_i^*\big) + \Delta_R(i,k),
\tag{4.33}
$$

where we defined

$$
\begin{aligned}
\Delta_R(i,k) &:= -\frac{\ell_i^4}{\|\boldsymbol{g}_i\|^2}\bar{h}_{ik}h_{ii}(\boldsymbol{e}_i+\boldsymbol{h}_i)(\boldsymbol{e}_i+\boldsymbol{h}_i)^* \\
&\quad + \ell_i^2\|\boldsymbol{g}_i\|^{-1}\overline{h}_{ik}\big(\boldsymbol{h}_i\boldsymbol{e}_i^* + \boldsymbol{e}_i\boldsymbol{h}_i^* + 2\boldsymbol{h}_i\boldsymbol{h}_i^*\big).
\end{aligned}
\tag{4.34}
$$

By (4.33) and the fact that $\widetilde{B}^{\langle i\rangle}$ is independent of g_{ik}, we obtain that

$$
\frac{\partial G}{\partial g_{ik}} = \frac{\ell_i^2}{\|\boldsymbol{g}_i\|}GA\big(\boldsymbol{e}_k\big(\boldsymbol{e}_i^* + \boldsymbol{h}_i^*\big)\widetilde{B}^{\langle i\rangle}R_i + R_i\widetilde{B}^{\langle i\rangle}\boldsymbol{e}_k\big(\boldsymbol{e}_i^* + \boldsymbol{h}_i^*\big)\big)G + \Delta_G(i,k),
\tag{4.35}
$$

where

$$
\Delta_G(i,k) := -GA\big(\Delta_R(i,k)\widetilde{B}^{\langle i\rangle}R_i + R_i\widetilde{B}^{\langle i\rangle}\Delta_R(i,k)\big)G.
\tag{4.36}
$$

We see from (4.35) and (I.8) of [18] that

(4.37)
$$\begin{aligned}
&\frac{1}{N}\sum_k^{(i)} \boldsymbol{e}_k^* \widetilde{B}^{\langle i\rangle}\frac{\partial G}{\partial g_{ik}}\boldsymbol{e}_i \\
&\quad= \frac{\ell_i^2}{\|\boldsymbol{g}_i\|}\frac{1}{N}\sum_k^{(i)} \boldsymbol{e}_k^* \widetilde{B}^{\langle i\rangle} GA\big(\boldsymbol{e}_k(\boldsymbol{e}_i^*+\boldsymbol{h}_i^*)\widetilde{B}^{\langle i\rangle}R_i + R_i\widetilde{B}^{\langle i\rangle}\boldsymbol{e}_k(\boldsymbol{e}_i^*+\boldsymbol{h}_i^*)\big)G\boldsymbol{e}_i + \mathrm{O}_\prec(\Pi_i^2) \\
&\quad= \frac{\ell_i^2}{\|\boldsymbol{g}_i\|}\frac{1}{N}\sum_k^{(i)} [a_k \boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_k(-\boldsymbol{h}_i^*\widetilde{B}-\boldsymbol{e}_i^*\widetilde{B})G\boldsymbol{e}_i + \boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}GAR_i\widetilde{B}^{\langle i\rangle}\boldsymbol{e}_k(G_{ii}+\boldsymbol{h}_i^*G\boldsymbol{e}_i)] \\
&\qquad+ \mathrm{O}_\prec(\Pi_i^2) \\
&\quad= \frac{\ell_i^2}{\|\boldsymbol{g}_i\|}\frac{1}{N}\sum_k^{(i)} a_k\big(\widetilde{B}^{\langle i\rangle}G\big)_{kk}(-b_iT_i-(\widetilde{B}G)_{ii}) \\
&\qquad+ \frac{\ell_i^2}{\|\boldsymbol{g}_i\|}\frac{1}{N}\sum_k^{(i)} (\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}GAR_i\widetilde{B}^{\langle i\rangle}\boldsymbol{e}_k)(G_{ii}+T_i) + \mathrm{O}_\prec(\Pi_i^2),
\end{aligned}$$

where in the second equality we used (3.6) and in the third equality we used (3.6) and (3.9). Moreover, by (4.2) and the assumption that $\{a_i\}$ and $\{b_i\}$ are bounded, we readily see that

$$\mathrm{tr}\big(A\widetilde{B}^{\langle i\rangle}G\big) - \frac{1}{N}\sum_k^{(i)} a_k\big(\widetilde{B}^{\langle i\rangle}G\big)_{kk} = \frac{1}{N}a_ib_iG_{ii} \prec \frac{1}{N}, \tag{4.38}$$

and

$$\mathrm{tr}\big(\widetilde{B}^{\langle i\rangle}GAR_i\widetilde{B}^{\langle i\rangle}\big) - \frac{1}{N}\sum_k^{(i)} \boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}GAR_i\widetilde{B}^{\langle i\rangle}\boldsymbol{e}_k = -\frac{b_i}{N}\boldsymbol{e}_i^*GA\widetilde{B}\boldsymbol{h}_i \prec \frac{1}{N}. \tag{4.39}$$

We claim that we can replace $\widetilde{B}^{\langle i\rangle}$ by $\widetilde{B}$ in (4.38) and (4.39) without changing the error bound in (4.37). In fact, by the definition of $\widetilde{B}^{\langle i\rangle}$, we have that

$$\begin{aligned}
\mathrm{tr}(A\widetilde{B}G) - \mathrm{tr}\big(A\widetilde{B}^{\langle i\rangle}G\big) &= \mathrm{tr}(A\widetilde{B}G) - \mathrm{tr}(AR_i\widetilde{B}R_iG) \\
&= \mathrm{tr}\big(A\boldsymbol{r}_i\boldsymbol{r}_i^*\widetilde{B}G\big) + \mathrm{tr}\big(A\widetilde{B}\boldsymbol{r}_i\boldsymbol{r}_i^*G\big) - \mathrm{tr}\big(A\boldsymbol{r}_i\boldsymbol{r}_i^*\widetilde{B}\boldsymbol{r}_i\boldsymbol{r}_i^*G\big) \\
&= \frac{1}{N}\boldsymbol{r}_i^*\widetilde{B}GA\boldsymbol{r}_i + \frac{1}{N}\boldsymbol{r}_i^*GA\widetilde{B}\boldsymbol{r}_i - \frac{1}{N}\boldsymbol{r}_i^*\widetilde{B}\boldsymbol{r}_i\boldsymbol{r}_i^*GA\boldsymbol{r}_i.
\end{aligned} \tag{4.40}$$

Recall that $\boldsymbol{r}_i = \ell_i(\boldsymbol{e}_i+\boldsymbol{h}_i)$. Then we have

$$\begin{aligned}
\Big|\frac{1}{N}\boldsymbol{r}_i^*\widetilde{B}GA\boldsymbol{r}_i\Big| &\lesssim \frac{1}{N}\big(\|\sqrt{A}G^*\widetilde{B}\boldsymbol{e}_i\|\,\|A^{1/2}\| + \|G^*\widetilde{B}\boldsymbol{h}_i\|\big) \\
&\lesssim \frac{1}{N}\big(\boldsymbol{e}_i\widetilde{B}GAG^*\widetilde{B}\boldsymbol{e}_i + b_i^2\boldsymbol{h}_i^*GG^*\boldsymbol{h}_i\big)^{1/2},
\end{aligned} \tag{4.41}$$

where in the second inequality we used the fact that $\|A\|$ is bounded. Moreover, using (B.2) of [18], (2.5) and (2.2), it is easy to see that

$$\boldsymbol{e}_i^*\widetilde{B}GAG^*\widetilde{B}\boldsymbol{e}_i = \frac{\boldsymbol{e}_i^*\sqrt{A}\widetilde{B}\sqrt{A}\widetilde{G}\widetilde{G}^*\sqrt{A}\widetilde{B}\sqrt{A}\boldsymbol{e}_i}{a_i} = \frac{\boldsymbol{e}_i^*\widetilde{H}\widetilde{G}\widetilde{G}^*\widetilde{H}\boldsymbol{e}_i}{a_i} \le \|\widetilde{H}\|^2\frac{\mathrm{Im}\,\widetilde{G}_{ii}}{a_i\eta}, \tag{4.42}$$

where in the last inequality we used the fact that $\widetilde{G}$ is Hermitian and the Ward identity

$$\sum_{j=1}^{N} |\widetilde{G}_{ij}|^2 = (\widetilde{G}\widetilde{G}^*)_{ii} = \frac{\operatorname{Im}\widetilde{G}_{ii}}{\eta}.$$

Similarly, we can show that

$$b_i^2 \boldsymbol{h}_i^* G G^* \boldsymbol{h}_i = b_i^2 \boldsymbol{e}_i \mathcal{G}\mathcal{G}^* \boldsymbol{e}_i = b_i \boldsymbol{e}_i \widetilde{\mathcal{G}} B \widetilde{\mathcal{G}}^* \boldsymbol{e}_i \le b_i \|B\| \frac{\operatorname{Im}\widetilde{\mathcal{G}}_{ii}}{\eta}. \tag{4.43}$$

Since A, B, $\widetilde{H}$ are bounded, by (4.41), (4.42) and (4.43), we see that

$$\left|\frac{1}{N}\boldsymbol{r}_i^* \widetilde{B} G A \boldsymbol{r}_i\right| \lesssim \frac{1}{N}\left(\frac{\operatorname{Im}\widetilde{G}_{ii}}{\eta} + \frac{\operatorname{Im}\widetilde{\mathcal{G}}_{ii}}{\eta}\right)^{1/2}.$$

By an analogous discussion, we can control the other two terms of the RHS of (4.40) as

$$\left|\frac{1}{N}\boldsymbol{r}_i^* G A \widetilde{B} \boldsymbol{r}_i\right| \lesssim \frac{1}{N}\left(\frac{\operatorname{Im}\widetilde{G}_{ii}}{\eta} + \frac{\operatorname{Im}\widetilde{\mathcal{G}}_{ii}}{\eta}\right)^{1/2},$$

$$\left|\frac{1}{N}\boldsymbol{r}_i^* \widetilde{B} \boldsymbol{r}_i \boldsymbol{r}_i^* G A \boldsymbol{r}_i\right| \lesssim \frac{1}{N}\left(\frac{\operatorname{Im}\widetilde{G}_{ii}}{\eta} + \frac{\operatorname{Im}\widetilde{\mathcal{G}}_{ii}}{\eta}\right)^{1/2}.$$

Furthermore, from the spectral decomposition of $\widetilde{H}$ and $\widetilde{\mathcal{H}}$, it is clear that $\operatorname{Im}\widetilde{G}_{ii}/\eta \ge c$ and $\operatorname{Im}\widetilde{\mathcal{G}}_{ii}/\eta \ge c$ for some fixed constant $c > 0$. This shows that for some constant $C > 0$,

$$\frac{1}{N}\left(\frac{\operatorname{Im}\widetilde{G}_{ii} + \operatorname{Im}\widetilde{\mathcal{G}}_{ii}}{\eta}\right)^{1/2} \le \frac{C}{N}\frac{\operatorname{Im}\widetilde{G}_{ii} + \operatorname{Im}\widetilde{\mathcal{G}}_{ii}}{\eta} = C\Pi_i^2.$$

Together with (4.40), we arrive at

$$\operatorname{tr}\big(A\widetilde{B}^{\langle i\rangle} G\big) = \operatorname{tr}(A\widetilde{B}G) + \mathrm{O}_{\prec}\big(\Pi_i^2\big). \tag{4.44}$$

By a discussion similar to (4.44), we can get

$$\operatorname{tr}\big(\widetilde{B}^{\langle i\rangle} G A R_i \widetilde{B}^{\langle i\rangle}\big) = \operatorname{tr}\left(\widetilde{B}GA\widetilde{B}\right) + \mathrm{O}_{\prec}\big(\Pi_i^2\big). \tag{4.45}$$

Therefore, by (4.37), (4.44) and (4.45), we conclude that

$$\begin{aligned}&\frac{1}{N}\sum_{k}^{(i)} \boldsymbol{e}_k^* \widetilde{B}^{\langle i\rangle}\frac{\partial G}{\partial g_{ik}}\boldsymbol{e}_i \\ &\quad= \frac{\ell_i^2}{\|\boldsymbol{g}_i\|}\big(\operatorname{tr}(A\widetilde{B}G)\big(-b_i T_i - (\widetilde{B}G)_{ii}\big) + \operatorname{tr}(\widetilde{B}GA\widetilde{B})(G_{ii} + T_i)\big) + \mathrm{O}_{\prec}\big(\Pi_i^2\big).\end{aligned} \tag{4.46}$$

Note that compared to the expansion (4.29), the coefficient in front of $\operatorname{tr}(A\widetilde{B}G)$ is still different. We need to further explore the hidden relation. By a discussion similar to (4.46), we have that

$$\begin{aligned}&\frac{1}{N}\sum_{k}^{(i)} \boldsymbol{e}_k^* \frac{\partial G}{\partial g_{ik}}\boldsymbol{e}_i \\ &\quad= \frac{\ell_i^2}{\|\boldsymbol{g}_i\|}\big(\operatorname{tr}(GA)\big(-b_i T_i - (\widetilde{B}G)_{ii}\big) + \operatorname{tr}(GA\widetilde{B})(G_{ii} + T_i)\big) + \mathrm{O}_{\prec}\big(\Pi_i^2\big).\end{aligned} \tag{4.47}$$

In light of (4.46), (4.47) and (4.29), it suffices to control

$$\operatorname{tr}(GA)\frac{1}{N}\sum_{k}^{(i)} \boldsymbol{e}_k^* \widetilde{B}^{\langle i\rangle}\frac{\partial G}{\partial g_{ik}}\boldsymbol{e}_i - \operatorname{tr}(A\widetilde{B}G)\frac{1}{N}\sum_{k}^{(i)} \boldsymbol{e}_k^* \frac{\partial G}{\partial g_{ik}}\boldsymbol{e}_i.$$

Combining (4.46) and (4.47), we have that

$$\begin{aligned}
&\operatorname{tr}(GA)\frac{1}{N}\sum_{k}^{(i)}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}\frac{\partial G}{\partial g_{ik}}\boldsymbol{e}_i-\operatorname{tr}(A\widetilde{B}G)\frac{1}{N}\sum_{k}^{(i)}\boldsymbol{e}_k^*\frac{\partial G}{\partial g_{ik}}\boldsymbol{e}_i\\
&=\frac{\ell_i^2}{\|\boldsymbol{g}_i\|}(G_{ii}+T_i)\big(\operatorname{tr}(GA)\operatorname{tr}(\widetilde{B}GA\widetilde{B})-\operatorname{tr}(GA\widetilde{B})\operatorname{tr}(GA\widetilde{B})\big)+\mathrm{O}_{\prec}\big(\Pi_i^2\big)\\
&=\frac{\ell_i^2}{\|\boldsymbol{g}_i\|}(G_{ii}+T_i)\big(-\Upsilon-\operatorname{tr}(A\widetilde{B}G)+\operatorname{tr}(GA)\big)+\mathrm{O}_{\prec}\big(\Pi_i^2\big),
\end{aligned}\tag{4.48}$$

where in the second equality we employed the definition of Υ in (3.12). Denote

$$\begin{aligned}
\mathsf{e}_{i2}:=&\bigg(\frac{\ell_i^2}{\|\boldsymbol{g}_i\|}-\|\boldsymbol{g}_i\|\bigg)\big(-G_{ii}\operatorname{tr}(A\widetilde{B}G)-(G_{ii}+T_i)\big(\Upsilon-\operatorname{tr}(GA)\big)\big)\\
&+\operatorname{tr}(A\widetilde{B}G)\bigg(\|\boldsymbol{g}_i\|\mathring{T}_i-\frac{\ell_i^2}{\|\boldsymbol{g}_i\|}T_i\bigg).
\end{aligned}\tag{4.49}$$

By a discussion similar to (4.28), we can conclude that

$$|\mathsf{e}_{i2}|\prec N^{-1/2}.\tag{4.50}$$

Moreover, by a simple algebraic calculation using (4.48) and (4.49), we find that

$$\begin{aligned}
&\operatorname{tr}(GA)\frac{1}{N}\sum_{k}^{(i)}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}\frac{\partial G}{\partial g_{ik}}\boldsymbol{e}_i\\
&=\|\boldsymbol{g}_i\|\big(-G_{ii}\operatorname{tr}(A\widetilde{B}G)-(G_{ii}+T_i)\big(\Upsilon-\operatorname{tr}(GA)\big)\big)\\
&\quad+\operatorname{tr}(A\widetilde{B}G)\bigg(\frac{1}{N}\sum_{k}^{(i)}\boldsymbol{e}_k^*\frac{\partial G}{\partial g_{ik}}\boldsymbol{e}_i-\|\boldsymbol{g}_i\|\mathring{T}_i\bigg)+\mathsf{e}_{i2}+\mathrm{O}_{\prec}\big(\Pi_i^2\big).
\end{aligned}\tag{4.51}$$

With (4.51) and (4.32), we can now come back to discussing (4.29). More specifically, inserting (4.51) into (4.32) and then (4.29), we have that

$$\begin{aligned}
\mathbb{E}\big[\mathfrak{X}_i^{(p,p)}\big]=&\mathbb{E}\bigg[\bigg(\frac{1}{N}\sum_{k}^{(i)}\frac{1}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k^*\frac{\partial G}{\partial g_{ik}}\boldsymbol{e}_i-\mathring{T}_i\bigg)\operatorname{tr}(A\widetilde{B}G)\mathfrak{X}_i^{(p-1,p)}\bigg]\\
&+\frac{1}{N}\sum_{k}^{(i)}\mathbb{E}\bigg[\frac{\partial\|\boldsymbol{g}_i\|^{-1}}{\partial g_{ik}}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i\operatorname{tr}(GA)\mathfrak{X}_i^{(p-1,p)}\bigg]\\
&+\frac{1}{N}\sum_{k}^{(i)}\mathbb{E}\bigg[\frac{\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i}{\|\boldsymbol{g}_i\|}\operatorname{tr}\bigg(\frac{\partial G}{\partial g_{ik}}A\bigg)\mathfrak{X}_i^{(p-1,p)}\bigg]\\
&+\frac{p-1}{N}\sum_{k}^{(i)}\mathbb{E}\bigg[\frac{1}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i\operatorname{tr}(GA)\frac{\partial P_i}{\partial g_{ik}}\mathfrak{X}_i^{(p-2,p)}\bigg]\\
&+\frac{p}{N}\sum_{k}^{(i)}\mathbb{E}\bigg[\frac{1}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i\operatorname{tr}(GA)\frac{\partial\overline{P}_i}{\partial g_{ik}}\mathfrak{X}_i^{(p-1,p-1)}\bigg]\\
&+\mathbb{E}\bigg[\bigg(\mathsf{e}_{i1}\operatorname{tr}(GA)+\frac{1}{\|\boldsymbol{g}_i\|}\mathsf{e}_{i2}+\mathrm{O}_{\prec}\big(\Pi_i^2\big)\bigg)\mathfrak{X}_i^{(p-1,p)}\bigg].
\end{aligned}\tag{4.52}$$

We do one more expansion for the first term of the above equation. Recall the definitions in (3.9). Applying the technique of integration by parts, that is, (4.31), we get that

$$\begin{aligned}
&\mathbb{E}\bigg[\bigg(\mathring{T}_i - \frac{1}{N}\sum_k^{(i)} \frac{1}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k^* \frac{\partial G}{\partial g_{ik}}\boldsymbol{e}_i\bigg)\operatorname{tr}(A\widetilde{B}G)\mathfrak{X}_i^{(p-1,p)}\bigg] \\
&\quad = \frac{1}{N}\sum_k^{(i)} \mathbb{E}\bigg[\frac{\partial \|\boldsymbol{g}_i\|^{-1}}{\partial g_{ik}} G_{ki}\operatorname{tr}(A\widetilde{B}G)\mathfrak{X}_i^{(p-1,p)}\bigg] \\
&\quad\quad + \frac{1}{N}\sum_k^{(i)} \mathbb{E}\bigg[\frac{1}{\|\boldsymbol{g}_i\|} G_{ki}\operatorname{tr}\bigg(A\widetilde{B}\frac{\partial G}{\partial g_{ik}}\bigg)\mathfrak{X}_i^{(p-1,p)}\bigg] \\
&\quad\quad + \frac{p-1}{N}\sum_k^{(i)} \mathbb{E}\bigg[\frac{1}{\|\boldsymbol{g}_i\|} G_{ki}\operatorname{tr}(A\widetilde{B}G)\frac{\partial P_i}{\partial g_{ik}}\mathfrak{X}_i^{(p-2,p)}\bigg] \\
&\quad\quad + \frac{p}{N}\sum_k^{(i)} \mathbb{E}\bigg[\frac{1}{\|\boldsymbol{g}_i\|} G_{ki}\operatorname{tr}(A\widetilde{B}G)\frac{\partial \overline{P}_i}{\partial g_{ik}}\mathfrak{X}_i^{(p-1,p-1)}\bigg].
\end{aligned} \tag{4.53}$$

Combining (4.53) and (4.52), we can rewrite

$$\mathbb{E}[\mathfrak{X}_i^{(p,p)}] = \mathbb{E}[\mathfrak{C}_1\mathfrak{X}_i^{(p-1,p)}] + \mathbb{E}[\mathfrak{C}_2\mathfrak{X}_i^{(p-2,p)}] + \mathbb{E}[\mathfrak{C}_3\mathfrak{X}_i^{(p-1,p-1)}], \tag{4.54}$$

where the coefficients $\mathfrak{C}_k, k = 1, 2, 3$ are defined as

$$\begin{aligned}
\mathfrak{C}_1 := \frac{1}{N}\sum_k^{(i)}\bigg(&\frac{\partial \|\boldsymbol{g}_i\|^{-1}}{\partial g_{ik}} G_{ki}\operatorname{tr}(A\widetilde{B}G) + \frac{1}{\|\boldsymbol{g}_i\|} G_{ki}\operatorname{tr}\bigg(A\widetilde{B}\frac{\partial G}{\partial g_{ik}}\bigg) \\
&+ \frac{\partial \|\boldsymbol{g}_i\|^{-1}}{\partial g_{ik}}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i\operatorname{tr}(GA) + \frac{\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i}{\|\boldsymbol{g}_i\|}\operatorname{tr}\bigg(\frac{\partial G}{\partial g_{ik}}A\bigg) \\
&+ \bigg(\mathsf{e}_{i1}\operatorname{tr}(GA) + \frac{1}{\|\boldsymbol{g}_i\|}\mathsf{e}_{i2} + \mathrm{O}_{\prec}(\Pi_i^2)\bigg)\bigg),
\end{aligned} \tag{4.55}$$

$$\mathfrak{C}_2 := \frac{p-1}{N}\sum_k^{(i)}\bigg(\frac{1}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i\operatorname{tr}(GA)\frac{\partial P_i}{\partial g_{ik}} + \frac{1}{\|\boldsymbol{g}_i\|}G_{ki}\operatorname{tr}(A\widetilde{B}G)\frac{\partial P_i}{\partial g_{ik}}\bigg), \tag{4.56}$$

$$\mathfrak{C}_3 := \frac{p}{N}\sum_k^{(i)}\bigg(\frac{1}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G\boldsymbol{e}_i\operatorname{tr}(GA)\frac{\partial \overline{P}_i}{\partial g_{ik}} + \frac{1}{\|\boldsymbol{g}_i\|}G_{ki}\operatorname{tr}(A\widetilde{B}G)\frac{\partial \overline{P}_i}{\partial g_{ik}}\bigg). \tag{4.57}$$

To conclude the proof of (4.19), it suffices to control the coefficients $\mathfrak{C}_k, k = 1, 2, 3$. For $\mathfrak{C}_1$, by Lemma I.4 of [18], we find that

$$\mathfrak{C}_1 \prec N^{-1/2} + \Pi_i^2. \tag{4.58}$$

For $\mathfrak{C}_2$, by (I.5) and Lemma I.4 of [18], we find that

$$\mathfrak{C}_2 \prec \Pi_i^2. \tag{4.59}$$

Similarly, we can show that

$$\mathfrak{C}_3 \prec \Pi_i^2. \tag{4.60}$$

Using (4.3), we complete the proof of (4.19) using (4.58), (4.59), (4.60) and (4.54).

Finally, due to similarity, we only briefly discuss the proof of (4.20). Using the definition of K_i in (3.11) and the fact that $T_i - \mathring{T}_i = h_{ii}G_{ii} \prec N^{-1/2}$, we find that

$$\begin{aligned}\mathbb{E}[\mathfrak{Y}_i^{(p,p)}] &= \mathbb{E}\big[\big(\mathring{T}_i + \operatorname{tr}(GA)(b_iT_i + (\widetilde{B}G)_{ii}) - \operatorname{tr}(GA\widetilde{B}G)(G_{ii}+T_i)\big)\mathfrak{Y}_i^{(p-1,p)}\big] \\ &\quad + \mathbb{E}[\mathrm{O}_\prec(N^{-1/2})\mathfrak{Y}_i^{(p-1,p)}] \\ &= \sum_k^{(i)} \mathbb{E}\Big[\frac{\overline{g}_{ik}}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k^* G\boldsymbol{e}_i \mathfrak{Y}_i^{(p-1,p)}\Big] \\ &\quad + \mathbb{E}\big[\big(\operatorname{tr}(GA)(b_iT_i + (\widetilde{B}G)_{ii}) - \operatorname{tr}(GA\widetilde{B}G)(G_{ii}+T_i)\big)\mathfrak{Y}_i^{(p-1,p)}\big] \\ &\quad + \mathbb{E}[\mathrm{O}_\prec(N^{-1/2})\mathfrak{Y}_i^{(p-1,p)}],\end{aligned} \tag{4.61}$$

where in the second equality we used the definition in (3.9). Applying (4.31) to the first term of the RHS of the above equation, we obtain

$$\begin{aligned}\sum_k^{(i)} \mathbb{E}\Big[\frac{\overline{g}_{ik}}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k^* G\boldsymbol{e}_i \mathfrak{Y}_i^{(p-1,p)}\Big] &= \frac{1}{N}\sum_k^{(i)} \mathbb{E}\Big[\frac{\partial \|\boldsymbol{g}_i\|^{-1}}{\partial g_{ik}}\boldsymbol{e}_k^* G\boldsymbol{e}_i \mathfrak{Y}_i^{(p-1,p)}\Big] \\ &\quad + \frac{1}{N}\sum_k^{(i)} \mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k^* \frac{\partial G}{\partial g_{ik}}\boldsymbol{e}_i \mathfrak{Y}_i^{(p-1,p)}\Big] \\ &\quad + \frac{p-1}{N}\sum_k^{(i)} \mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k^* G\boldsymbol{e}_i \frac{\partial K_i}{\partial g_{ik}} \mathfrak{Y}_i^{(p-2,p)}\Big] \\ &\quad + \frac{p}{N}\sum_k^{(i)} \mathbb{E}\Big[\frac{1}{\|\boldsymbol{g}_i\|}\boldsymbol{e}_k^* G\boldsymbol{e}_i \frac{\partial \overline{K}_i}{\partial g_{ik}} \mathfrak{Y}_i^{(p-1,p-1)}\Big].\end{aligned} \tag{4.62}$$

Inserting (4.47) into (4.62) and then (4.61), by a discussion similar to the cancellation in (4.52) and error controls in (4.28) and (4.50), we conclude that

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

Using (I.6), Lemma I.4 of [18] and a discussion similar to (4.58), (4.59) and (4.60), we can finish the proof of (4.20). □

Acknowledgments. The authors would like to thank the Editor, Associate Editor and an anonymous referee for their many critical suggestions which have significantly improved the paper. We also want to thank Zhigang Bao and Ji Oon Lee for many helpful discussions and comments.

Funding. The first author is partially supported by NSF Grant DMS-2113489 and grateful for the AMS-SIMONS travel grant (2020–2023). The second author is supported by the ERC Advanced Grant "RMTBeyond" No. 101020331.

SUPPLEMENTARY MATERIAL

Supplement to "Local laws for multiplication of random matrices" (DOI: 10.1214/22-AAP1882SUPP; .pdf). In the supplementary file, we provide the technical proofs for our main results in Section 2. We also collect and prove some auxiliary lemmas.

REFERENCES

[1] ABBE, E. (2017). Community detection and stochastic block models: Recent developments. *J. Mach. Learn. Res.* **18** Paper No. 177, 86 pp. MR3827065

[2] BAO, Z., ERDŐS, L. and SCHNELLI, K. (2016). Local stability of the free additive convolution. *J. Funct. Anal.* **271** 672–719. MR3506962 https://doi.org/10.1016/j.jfa.2016.04.006

[3] BAO, Z., ERDŐS, L. and SCHNELLI, K. (2017). Local law of addition of random matrices on optimal scale. *Comm. Math. Phys.* **349** 947–990. MR3602820 https://doi.org/10.1007/s00220-016-2805-6

[4] BAO, Z., ERDŐS, L. and SCHNELLI, K. (2017). Convergence rate for spectral distribution of addition of random matrices. *Adv. Math.* **319** 251–291. MR3695875 https://doi.org/10.1016/j.aim.2017.08.028

[5] BAO, Z., ERDŐS, L. and SCHNELLI, K. (2019). Local single ring theorem on optimal scale. *Ann. Probab.* **47** 1270–1334. MR3945747 https://doi.org/10.1214/18-AOP1284

[6] BAO, Z., ERDŐS, L. and SCHNELLI, K. (2020). On the support of the free additive convolution. *J. Anal. Math.* **142** 323–348. MR4205272 https://doi.org/10.1007/s11854-020-0135-2

[7] BAO, Z., ERDŐS, L. and SCHNELLI, K. (2020). Spectral rigidity for addition of random matrices at the regular edge. *J. Funct. Anal.* **279** 108639, 94 pp. MR4102163 https://doi.org/10.1016/j.jfa.2020.108639

[8] BAO, Z., PAN, G. and ZHOU, W. (2015). Universality for the largest eigenvalue of sample covariance matrices with general population. *Ann. Statist.* **43** 382–421. MR3311864 https://doi.org/10.1214/14-AOS1281

[9] BELINSCHI, S. T. (2006). A note on regularity for free convolutions. *Ann. Inst. Henri Poincaré Probab. Stat.* **42** 635–648. MR2259979 https://doi.org/10.1016/j.anihpb.2005.05.004

[10] BELINSCHI, S. T. and BERCOVICI, H. (2007). A new approach to subordination results in free probability. *J. Anal. Math.* **101** 357–365. MR2346550 https://doi.org/10.1007/s11854-007-0013-1

[11] BELINSCHI, S. T., BERCOVICI, H., CAPITAINE, M. and FÉVRIER, M. (2017). Outliers in the spectrum of large deformed unitarily invariant models. *Ann. Probab.* **45** 3571–3625. MR3729610 https://doi.org/10.1214/16-AOP1144

[12] BELINSCHI, S. T., MAI, T. and SPEICHER, R. (2017). Analytic subordination theory of operator-valued free additive convolution and the solution of a general random matrix problem. *J. Reine Angew. Math.* **732** 21–53. MR3717087 https://doi.org/10.1515/crelle-2014-0138

[13] BIANE, P. (1997). On the free convolution with a semi-circular distribution. *Indiana Univ. Math. J.* **46** 705–718. MR1488333 https://doi.org/10.1512/iumj.1997.46.1467

[14] BUN, J., ALLEZ, R., BOUCHAUD, J.-P. and POTTERS, M. (2016). Rotational invariant estimator for general noisy matrices. *IEEE Trans. Inf. Theory* **62** 7475–7490. MR3599095 https://doi.org/10.1109/TIT.2016.2616132

[15] BUN, J., BOUCHAUD, J.-P. and POTTERS, M. (2017). Cleaning large correlation matrices: Tools from random matrix theory. *Phys. Rep.* **666** 1–109. MR3590056 https://doi.org/10.1016/j.physrep.2016.10.005

[16] CHISTYAKOV, G. P. and GÖTZE, F. (2011). The arithmetic of distributions in free probability theory. *Cent. Eur. J. Math.* **9** 997–1050. MR2824443 https://doi.org/10.2478/s11533-011-0049-4

[17] DIACONIS, P. and SHAHSHAHANI, M. (1987). The subgroup algorithm for generating uniform random variables. *Probab. Engrg. Inform. Sci.* **1** 15–32.

[18] DING, X. and JI, H. C. (2023). Supplement to "Local laws for multiplication of random matrices." https://doi.org/10.1214/22-AAP1882SUPP

[19] DING, X. and WU, H.-T. (2021). On the spectral property of kernel-based sensor fusion algorithms of high dimensional data. *IEEE Trans. Inf. Theory* **67** 640–670. MR4231977 https://doi.org/10.1109/TIT.2020.3026255

[20] DING, X. and WU, H.-T. (2021). How do kernel-based sensor fusion algorithms behave under high dimensional noise? Preprint. Available at arXiv:2111.10940.

[21] DING, X. and YANG, F. (2018). A necessary and sufficient condition for edge universality at the largest singular values of covariance matrices. *Ann. Appl. Probab.* **28** 1679–1738. MR3809475 https://doi.org/10.1214/17-AAP1341

[22] DING, X. and YANG, F. (2021). Spiked separable covariance matrices and principal components. *Ann. Statist.* **49** 1113–1138. MR4255121 https://doi.org/10.1214/20-aos1995

[23] DING, X. and YANG, F. (2022). Tracy–Widom distribution for heterogeneous Gram matrices with applications in signal detection. *IEEE Trans. Inf. Theory* **68** 6682–6715. https://doi.org/10.1109/TIT.2022.3176784

[24] DOBRIBAN, E. and LIU, S. (2019). Asymptotics for sketching in least squares regression. In *Conference on Neural Information Processing Systems* (*NIPS*).

[25] DONOHO, D. L., GAVISH, M. and MONTANARI, A. (2013). The phase transition of matrix recovery from Gaussian measurements matches the minimax MSE of matrix denoising. *Proc. Natl. Acad. Sci. USA* **110** 8405–8410. MR3082268 https://doi.org/10.1073/pnas.1306110110

[26] ERDŐS, L., KNOWLES, A. and YAU, H.-T. (2013). Averaging fluctuations in resolvents of random band matrices. *Ann. Henri Poincaré* **14** 1837–1926. MR3119922 https://doi.org/10.1007/s00023-013-0235-y

[27] ERDŐS, L., KRÜGER, T. and NEMISH, Y. (2020). Local laws for polynomials of Wigner matrices. *J. Funct. Anal.* **278** 108507, 59 pp. MR4078529 https://doi.org/10.1016/j.jfa.2020.108507

[28] ERDŐS, L. and YAU, H.-T. (2017). *A Dynamical Approach to Random Matrix Theory*. *Courant Lecture Notes in Mathematics* **28**. Courant Institute of Mathematical Sciences, New York; Amer. Math. Soc., Providence, RI. MR3699468

[29] HO, C.-W. (2022). A local limit theorem and delocalization of eigenvectors for polynomials in two matrices. *Int. Math. Res. Not. IMRN* **2022** 1734–1769. MR4373224 https://doi.org/10.1093/imrn/rnaa116

[30] JAVANMARD, A., MONTANARI, A. and RICCI-TERSENGHI, F. (2016). Phase transitions in semidefinite relaxations. *Proc. Natl. Acad. Sci. USA* **113** E2218–E2223. MR3494080 https://doi.org/10.1073/pnas.1523097113

[31] JI, H. C. (2021). Regularity properties of free multiplicative convolution on the positive line. *Int. Math. Res. Not. IMRN* **2021** 4522–4563. MR4230404 https://doi.org/10.1093/imrn/rnaa152

[32] KARGIN, V. (2015). Subordination for the sum of two random matrices. *Ann. Probab.* **43** 2119–2150. MR3353823 https://doi.org/10.1214/14-AOP929

[33] KNOWLES, A. and YIN, J. (2017). Anisotropic local laws for random matrices. *Probab. Theory Related Fields* **169** 257–352. MR3704770 https://doi.org/10.1007/s00440-016-0730-4

[34] KWAK, J., LEE, J. O. and PARK, J. (2021). Extremal eigenvalues of sample covariance matrices with general population. *Bernoulli* **27** 2740–2765. MR4303902 https://doi.org/10.3150/21-BEJ1329

[35] LACOTTE, J. and PILANCI, M. (2020). Effective dimension adaptive sketching methods for faster regularized least-squares optimization. In *Conference on Neural Information Processing Systems* (*NIPS*).

[36] LEE, J. O. and SCHNELLI, K. (2016). Tracy–Widom distribution for the largest eigenvalue of real sample covariance matrices with general population. *Ann. Appl. Probab.* **26** 3786–3839. MR3582818 https://doi.org/10.1214/16-AAP1193

[37] LELARGE, M. and MIOLANE, L. (2019). Fundamental limits of symmetric low-rank matrix estimation. *Probab. Theory Related Fields* **173** 859–929. MR3936148 https://doi.org/10.1007/s00440-018-0845-x

[38] LIU, S. and DOBRIBAN, E. (2020). Ridge regression: Structure, cross-validation, and sketching. In *The* 8*th International Conference on Learning Representations* (*ICLR*).

[39] ONATSKI, A. (2009). Testing hypotheses about the numbers of factors in large factor models. *Econometrica* **77** 1447–1479. MR2561070 https://doi.org/10.3982/ECTA6964

[40] PASTUR, L. and VASILCHUK, V. (2000). On the law of addition of random matrices. *Comm. Math. Phys.* **214** 249–286. MR1796022 https://doi.org/10.1007/s002200000264

[41] PAUL, D. and SILVERSTEIN, J. W. (2009). No eigenvalues outside the support of the limiting empirical spectral distribution of a separable covariance matrix. *J. Multivariate Anal.* **100** 37–57. MR2460475 https://doi.org/10.1016/j.jmva.2008.03.010

[42] SCHWARTZMAN, A., MASCARENHAS, W. F. and TAYLOR, J. E. (2008). Inference for eigenvalues and eigenvectors of Gaussian symmetric matrices. *Ann. Statist.* **36** 2886–2919. MR2485016 https://doi.org/10.1214/08-AOS628

[43] TULINO, A. M., VERDÚ, S. et al. (2004). Random matrix theory and wireless communications. *Found. Trends Commun. Inf. Theory* **1** 1–182.

[44] VOICULESCU, D. (1987). Multiplication of certain noncommuting random variables. *J. Operator Theory* **18** 223–235. MR0915507

[45] VOICULESCU, D. (1991). Limit laws for random matrices and free products. *Invent. Math.* **104** 201–220. MR1094052 https://doi.org/10.1007/BF01245072

[46] VOICULESCU, D. V., DYKEMA, K. J. and NICA, A. (1992). *Free Random Variables*: *A Noncommutative Probability Approach to Free Products with Applications to Random Matrices*, *Operator Algebras and Harmonic Analysis on Free Groups*. *CRM Monograph Series* **1**. Amer. Math. Soc., Providence, RI. MR1217253 https://doi.org/10.1090/crmm/001

[47] YANG, F. (2019). Edge universality of separable covariance matrices. *Electron. J. Probab.* **24** Paper No. 123, 57 pp. MR4029426 https://doi.org/10.1214/19-ejp381

[48] YANG, F., LIU, S., DOBRIBAN, E. and WOODRUFF, D. P. (2021). How to reduce dimension with PCA and random projections? *IEEE Trans. Inf. Theory* **67** 8154–8189. MR4346082 https://doi.org/10.1109/tit.2021.3112821

[49] YAO, J., ZHENG, S. and BAI, Z. (2015). *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. *Cambridge Series in Statistical and Probabilistic Mathematics* **39**. Cambridge Univ. Press, New York. MR3468554 https://doi.org/10.1017/CBO9781107588080