PRINCIPAL COMPONENTS IN LINEAR MIXED MODELS WITH GENERAL BULK

ZHOU FAN, YI SUN, AND ZHICHAO WANG

Abstract. We study the outlier eigenvalues and eigenvectors in variance components estimates for high-dimensional mixed effects linear models using a free probability approach. We quantify the almost-sure limits of these eigenvalues and their eigenvector alignments, under a general bulk-plus-spikes assumption for the population covariances of the random effects, extending previous results in the identity-plus-spikes setting. Our analysis develops two tools in free probability and random matrix theory which are of independent interest—strong asymptotic freeness of GOE and deterministic matrices, and a method of proving deterministic anisotropic approximations to resolvents using free deterministic equivalents. Statistically, our results quantify bias and aliasing effects for the leading principal components of variance components estimates in modern high-dimensional applications.

1. Introduction

Principal components analysis (PCA) is a commonly used technique for identifying linear low-rank structure in high-dimensional data [Jol11]. For n independent samples in a comparable dimension p, when the entrywise noise variance in $\mathbb{R}^p$ is comparable in magnitude to the principal eigenvalues, it is now well-established that the principal components of the sample covariance matrix may be inaccurate for their population counterparts [JL09]. A body of work has quantified the behavior of PCA in this setting [Joh01, BBP05, BS06, Pau07, BGN12, BY12], connecting to the Marcenko-Pastur and Tracy-Widom laws of asymptotic random matrix theory [MP67, TW96]. We refer readers to the review articles [PA14, JP18] for more discussion and references to this and related lines of work.

Similar phenomena are to be expected in statistical models where samples are not independent, but instead exhibit a potentially complicated dependence structure [BJW05, Zha06, LAP15, WAP17]. However, the behavior of PCA in many such applications is less well-understood. The primary purpose of this work is to study one specific setting—that of mixed effects linear models [SCM09]—where dependence across observed samples arises via linear combinations of unobserved latent variables. We precisely characterize the almost sure limits of principal eigenvalues and eigenvectors of variance component estimates in these models, quantifying the high-dimensional bias and aliasing effects which arise in PCA in this context. This extends aspects of previous work [FJS18], which obtained such results under a restrictive assumption of isotropic noise, discussed below. This extension is important in practice, as such an assumption is conceptually illustrative but may be unrealistic as a model for real data.

Our techniques for studying this model are different from the direct analytic approach of [FJS18]. Instead, they generalize those of [FJ16] and are based on tools in free probability theory and its connection to random matrices [Voi91, MS17]. A second purpose of this work is to establish two general results in this area—strong asymptotic freeness of independent GOE and deterministic matrices, and a method of deriving anisotropic resolvent approximations using free deterministic equivalents [SV12]. Free probability techniques have recently been applied to study outliers in other random matrix models [BBCF17, BBC17] as well as more general questions about spectral behavior in other statistical problems, for example the analysis of autocovariance estimates for high-dimensional time series [BB16a, BB16b, BB17] and sketching methods for linear regression [DL18]. We believe that the tools we develop may be of broader interest to the analysis of structured random matrices arising in other statistics and engineering applications.

1.1. PCA in high-dimensional linear mixed models.

We study random and mixed effects linear models in a high-dimensional asymptotic framework. Extending the representation of [Rao72] to a multivariate

Z.F.: Department of Statistics and Data Science, Yale University, New Haven, CT 06511
Y.S.: Department of Mathematics, Columbia University, 2990 Broadway, New York, NY 10027
Z.W.: Department of Mathematics, Texas A&M University, College Station, TX 77840
E-mail addresses: zhou.fan@yale.edu, yisun@math.columbia.edu, wangzc@tamu.edu.

setting, such models take the form

$$Y = X\beta + U_1\alpha_1 + \ldots + U_k\alpha_k \in \mathbb{R}^{n\times p}$$

where Y contains n dependent samples in dimension p, each a combination of random effects constituting the rows of (unobserved) matrices $\alpha_1, \ldots, \alpha_k$. We study the behavior of PCA for classical MANOVA-type estimates of the *variance components* in these models, which are the population covariance matrices $\Sigma_1, \ldots, \Sigma_k$ for the random effects $\alpha_1, \ldots, \alpha_k$. The form of any one such estimate is the matrix

$$\widehat{\Sigma} = Y^{\mathsf{T}} B Y,$$

where $B \in \mathbb{R}^{n\times n}$ is a symmetric deterministic matrix satisfying $BX = 0$. Assuming that each Σ_r has a general spiked structure, our main results in this context, Theorems 2.6 and 2.7 below, characterize the first-order behavior of the principal eigenvalues and eigenvectors for the estimate $\widehat{\Sigma}$.

Mixed effects linear models are notably used within statistical genetics to model the variation of quantitative phenotypes across related samples of a population [LW98]. In this context, $U_1, \ldots, U_k$ may encode known pedigree of the samples as in twin/sibling studies or experimental breeding designs. Alternatively, they may capture relatedness via genotype values measured at a set of single-nucleotide polymorphisms (SNPs) [YLGV11, ZS12]. It has been recognized since the work of Fisher and Wright [Fis18, Wri35] that variance components in these models can provide a decomposition of the total variation in phenotypic traits into constituent genetic and non-genetic effects, thus yielding estimates of heritability. In high dimensions, the principal directions of variation in the genetic components may indicate phenotypic subspaces near which either responses to selection or random mutational drift are likely to be constrained [HB06, BM15, CMA+18]. Principal directions of variation in the environmental components may indicate effects of hidden experimental confounders, to be removed before performing downstream analyses [LS07, SPP+12].

High-dimensional asymptotic analysis for the spectral behavior of $\widehat{\Sigma}$ was initiated in [FJ16], which characterized the empirical eigenvalue distribution in the $n, p \to \infty$ limit. Individual outlier eigenvalues and eigenvectors were studied in greater detail in [FJS18], building on results of [FJ17], for a model with isotropic noise, meaning each population covariance Σ_r is a low-rank perturbation of $\sigma_r^2 \,\mathrm{Id}$. New qualitative phenomena were observed in [FJS18] which do not manifest in the setting of usual sample covariances for independent samples: Principal eigenvectors of an estimate of Σ_r may be biased towards those of a different component Σ_s, and the biases of principal eigenvalues and associated BBP-type phase transitions [BBP05, BS06, Pau07] depend collectively on the interaction between $\Sigma_1, \ldots, \Sigma_k$. Based on quantitative characterizations of these phenomena, a novel algorithm was developed in [FJS18] for removing these biases in PCA.

The isotropic noise assumption is restrictive in the modeling of real data. In this work, we generalize the first-order probabilistic results of [FJS18] beyond this assumption, by considering a general bulk-plus-spikes structure for each Σ_r described in Assumption 2.1 below. We show that the almost sure limits of principal eigenvalues and eigenvector alignments in [FJS18] extend naturally to this setting, involving quantities appearing in the fixed-point equations for the bulk law.

1.2. **Free probability results.** Our proofs use the connection between free probability and random matrices. Introducing appropriate representations of U_r, α_r, and B, detailed in Section 2.1, our matrix model $\widehat{\Sigma}$ may be written as

$$\widehat{\Sigma} = W + P, \qquad W = \sum_{r=1}^{k}\sum_{s=1}^{k} H_r^{\mathsf{T}} G_r^{\mathsf{T}} F_{rs} G_s H_s, \tag{1.1}$$

for deterministic matrices $\{H_1, \ldots, H_k\}$ and $\{F_{11}, F_{12}, \ldots, F_{kk}\}$, independent random Gaussian matrices $\{G_1, \ldots, G_k\}$, and a fixed-rank perturbation P (dependent on $G_1, \ldots, G_k$). The bulk eigenvalue distribution of W may be studied by introducing an asymptotic approximation

$$w = \sum_{r=1}^{k}\sum_{s=1}^{k} h_r^* g_r^* f_{rs} g_s h_s,$$

where h_r, g_r, f_{rs} belong to a von Neumann algebra $\mathcal{A}$ and are conditionally free (i.e. free with amalgamation) with respect to a trace $\tau : \mathcal{A} \to \mathbb{C}$ over a diagonal subalgebra $\mathcal{D}$ [BG09]. The spectral distribution of W may then be deduced via a computation of the τ-distribution of $w \in \mathcal{A}$. This approach was applied in [FJ16] to derive fixed-point equations, (2.5–2.7) below, that characterize the Stieltjes transform of a deterministic equivalent spectral law μ_0.

Such ideas connecting free probability to random matrices were introduced in the seminal work [Voi91] for polynomials of deterministic and GUE matrices, and subsequently extended to other matrix models in [Dyk93, Voi98, HP00, Col03, CC04, CŚ06]. Notably, we rely on results in [BG09] which establish conditional freeness of rectangular matrices embedded in a larger square space, and [SV12] which introduced the notion of a free deterministic equivalent where $\mathcal{A}$ may also be n-dependent. This latter construction is more natural for approximating matrix models with deterministic matrix components, removing the requirement of their (joint) convergence in spectral law and directly yielding a deterministic equivalent measure [HLN07].

Our study of outlier eigenvalues of $\widehat{\Sigma}$ in the presence of a fixed-rank perturbation P relies on the following two general results.

Strong asymptotic freeness of GOE and deterministic matrices. To characterize outlier eigenvalues of $\widehat{\Sigma}$ arising from P in (1.1), we first show that when $P = 0$, no eigenvalues of W separate from $\operatorname{supp}(\mu_0)$, where μ_0 is the deterministic equivalent measure for the bulk matrix W. We do this also using a free probability approach, showing for any fixed $\delta > 0$ the spectral inclusion

$$\operatorname{spec}(W) \subset \operatorname{spec}(w)_\delta$$

for all large n, where $\operatorname{spec}(w) = \operatorname{supp}(\mu_0)$ is the spectrum of w as a bounded operator in $\mathcal{A}$, and $\operatorname{spec}(w)_\delta$ is its δ-neighborhood.

We derive this as a consequence of a strong asymptotic freeness result for deterministic and GOE matrices, which parallels the GUE results in [Mal12]: Fix integers $p, q \geq 0$. Consider independent GOE matrices $X_1, \dots, X_p \in \mathbb{R}^{N\times N}$, deterministic matrices $Y_1, \dots, Y_q \in \mathbb{C}^{N\times N}$ with bounded operator norm, and a deterministic equivalent model $(\mathcal{A}, \tau)$ containing free semicircular elements $x_1, \dots, x_p$ and the matrices $Y_1, \dots, Y_q$. Theorem 3.1 below establishes

$$\operatorname{spec}(Q(X_1, \dots, X_p, Y_1, \dots, Y_q)) \subset \operatorname{spec}(Q(x_1, \dots, x_p, Y_1, \dots, Y_q))_\delta \tag{1.2}$$

for any fixed self-adjoint $*$-polynomial Q and $\delta > 0$, almost surely for all large N. If $Y_1, \dots, Y_q$ also converge both in joint tracial law and in operator norm to fixed limits $y_1, \dots, y_q$ of a fixed von Neumann algebra $\mathcal{A}$, then one may derive the more usual statement

$$\lim_{N\to\infty} \|Q(X_1, \dots, X_p, Y_1, \dots, Y_q)\| = \|Q(x_1, \dots, x_p, y_1, \dots, y_q)\|,$$

which we state as Theorem 3.2. In our application, we will apply the result directly in the form (1.2).

This is a strong freeness result extending the free approximation from the level of the normalized trace to that of the operator norm. The first such result was established in [HT05] for independent GUE matrices. This was extended to GOE matrices in [Sch05], complex Wigner matrices in [CDM07], GUE and deterministic matrices in [Mal12], and complex Wigner and deterministic matrices in [BC17]. The above result is an analogue of the results of [Mal12, BC17] in the real Gaussian setting. The proof follows closely the ideas of [HT05, Sch05, Mal12, BC17], and we discuss this further in Section 3.

Anisotropic resolvent laws from deterministic equivalents. To study outlier eigenvalues and eigenvectors produced by P, we follow the perturbative approach of [BGN11]. In particular, the outlier eigenvalues are the roots of an analytic equation $0 = \det \widehat{T}(z)$ for a fixed-dimensional matrix $\widehat{T}(z)$. We show $\widehat{T}(z) - T(z) \to 0$ entrywise for a deterministic approximation $T(z)$, uniformly over the spectral domain

$$U_\delta = \{z \in \mathbb{C} : \operatorname{dist}(z, \operatorname{supp}(\mu_0)) \geq \delta\}.$$

As is common to this method of analysis, the matrix $\widehat{T}(z)$ depends on terms of the form

$$u^* R(z) v, \qquad R(z) = (W - z \operatorname{Id})^{-1}$$

where $u, v \in \mathbb{C}^p$ are deterministic vectors related to the perturbation P, and $R(z)$ is the resolvent of the bulk matrix W in (1.1). In models where the rank-one matrix vu^* is "infinitesimally free" of W (for example, if W is rotationally invariant with respect to u, v) [Shl15, CHS18], such a term may be approximated as

$$u^* R(z) v \approx \langle u, v\rangle \cdot p^{-1} \operatorname{Tr} R(z) \approx \langle u, v\rangle \cdot m_0(z),$$

where m_0 is the Stieltjes transform of μ_0. This assumption does not hold in our setting because perturbative directions in one covariance Σ_r may be partially aligned with another covariance Σ_s. We show instead that

for a deterministic approximation $R_0(z) \in \mathbb{C}^{p\times p}$ to the resolvent, we have

$$\sup_{z\in U_\delta} |u^* R(z) v - u^* R_0(z) v| \to 0$$

for any deterministic unit vectors $u, v \in \mathbb{R}^p$.

Our analysis is general, and we consider the resolvent $R(z)$ of a matrix $W \in \mathbb{C}^{N\times N}$ which is an arbitrary self-adjoint $*$-polynomial of deterministic matrices $\{H_1, \ldots, H_p\}$ and random matrices $\{B_1, \ldots, B_q\}$. We assume that the latter are jointly orthogonally invariant in law, and hence free of the former. In this setting, Theorem 4.2 below establishes a deterministic approximation

$$R(z) \approx R_0(z)$$

in the above sense, where $R_0(z)$ is computable in the free deterministic equivalent framework of [SV12]. More concretely, let $(\mathcal{A}, \tau)$ be the von Neumann free product of $(\mathcal{A}_1, \tau_1) \equiv (\mathbb{C}^{N\times N}, N^{-1}\operatorname{Tr})$ containing $\{H_1, \ldots, H_p\}$ and $(\mathcal{A}_2, \tau_2)$ containing $\{b_1, \ldots, b_q\}$ which approximate $\{B_1, \ldots, B_q\}$ in joint law. Defining $w \in \mathcal{A}$ as the operator modeling W, the resolvent approximation is given by

$$R_0(z) = \tau^{\mathcal{H}}((w - z)^{-1})$$

where $\tau^{\mathcal{H}}$ is the τ-compatible conditional expectation onto the subalgebra generated by $H_1, \ldots, H_p$. Importantly, in the free deterministic equivalent (as opposed to purely asymptotic) framework, this subalgebra is contained in $\mathcal{A}_1 \equiv \mathbb{C}^{N\times N}$ so that $\tau^{\mathcal{H}}(a)$ is an $N \times N$ matrix for any $a \in \mathcal{A}$.

For z allowed to converge to $\operatorname{supp}(\mu_0)$, this type of result is an *anisotropic local law* [KY17]. This setting commonly arises in matrix models involving deterministic components. Our result is weaker in form than that of [KY17], as we consider only $z \in U_\delta$ with constant separation $\delta > 0$ from $\operatorname{supp}(\mu_0)$. However, our model for $\widehat{\Sigma}$ is more complicated than those studied in [KY17], and for which (to our knowledge) no such local result is currently available. We derive the anisotropic resolvent approximation using a simple free probability proof, which also applies directly to other matrix models. We mention that we expect such a result to hold in other settings where freeness arises, outside of orthogonal rotational invariance which we study in this work, although we do not pursue this direction here.

Acknowledgments. We thank Camille Male and Roland Speicher for helpful pointers to the strong asymptotic freeness literature. Y. S. was supported by a Junior Fellow award from the Simons Foundation and NSF Grant DMS-1701654.

2. Eigenvalues and eigenvectors in the mixed effects model

2.1. Definition of the model.

We study a multivariate mixed effects linear model

$$Y = X\beta + U_1\alpha_1 + \ldots + U_k\alpha_k \in \mathbb{R}^{n\times p}$$

with a p-dimensional response. Here, $X\beta$ denotes fixed effects and $U_1\alpha_1, \ldots, U_k\alpha_k$ random effects, where

- $X \in \mathbb{R}^{n\times m}$ is a known design matrix with unknown regression coefficients $\beta \in \mathbb{R}^{m\times p}$.
- For each $r = 1, \ldots, k$, the matrix $\alpha_r \in \mathbb{R}^{n_r\times p}$ is unobserved, and its rows constitute n_r independent realizations of a p-dimensional random effect.
- Each $U_r \in \mathbb{R}^{n\times n_r}$ is a known, deterministic incidence matrix specified by the model design.

A possible residual effect $\varepsilon \in \mathbb{R}^{n\times p}$, independent across samples, may be included by allowing the last random effect to be $\alpha_k = \varepsilon$ and $U_k = \mathrm{Id}$. This model encompasses many instances of common classification designs, discussed in greater detail in [FJ16, FJS18].

Our focus is on principal components analysis for the variance components of this model, which are the covariance matrices of the random effects. We assume that the random effects arise in the following way.

Assumption 2.1. The matrices $\alpha_1, \ldots, \alpha_k$ are independent. The rows of each α_r are independent, with the i^{th} row given by

$$\sum_{j=1}^{\ell_r} \gamma_j^{(r)} \xi_{ij}^{(r)} + \varepsilon_i^{(r)}.$$

Here $\gamma_1^{(r)}, \ldots, \gamma_{\ell_r}^{(r)} \in \mathbb{R}^p$ are ℓ_r deterministic directions of variation, $\xi_{ij}^{(r)} \in \mathbb{R}$ are independent and satisfy

$$\mathbb{E}[\xi_{ij}^{(r)}] = 0, \qquad \mathbb{E}[(\xi_{ij}^{(r)})^2] = 1, \qquad \mathbb{E}[|\xi_{ij}^{(r)}|^k] \le C_k$$

for all $k \geq 1$ and some constants $C_k > 0$, and

$$\varepsilon_i^{(r)} \sim \mathcal{N}(0, \mathring{\Sigma}_r)$$

for a noise covariance $\mathring{\Sigma}_r \in \mathbb{R}^{p \times p}$.

As a compromise between generality of the model and simplicity of the analysis, we impose a normality assumption on the noise $\varepsilon_i^{(r)}$ but not on the coefficients $\xi_{ij}^{(r)}$—a similar approach was used in [Nad08]. As $\xi_{ij}^{(r)}$ are not normal, we correspondingly will not assume that $\gamma_1^{(r)}, \dots, \gamma_{\ell_r}^{(r)}$ are orthogonal for each r. Introducing

$$\Gamma_r = \begin{pmatrix} - & \gamma_1^{(r)} & - \\ & \vdots & \\ - & \gamma_{\ell_r}^{(r)} & - \end{pmatrix} \in \mathbb{R}^{\ell_r \times p}, \qquad \Xi_r = \frac{1}{\sqrt{n_r}} (\xi_{ij}^{(r)})_{i,j} \in \mathbb{R}^{n_r \times \ell_r}, \qquad E_r = \begin{pmatrix} - & \varepsilon_1^{(r)} & - \\ & \vdots & \\ - & \varepsilon_n^{(r)} & - \end{pmatrix} \in \mathbb{R}^{n_r \times p},$$

the model for α_r is written concisely as

$$\alpha_r = \sqrt{n_r} \Xi_r \Gamma_r + E_r. \tag{2.1}$$

The rows of each α_r are then independent with mean 0 and covariance of the spiked form

$$\Sigma_r = \Gamma_r^{\mathsf{T}} \Gamma_r + \mathring{\Sigma}_r, \tag{2.2}$$

where $\Gamma_r^{\mathsf{T}} \Gamma_r$ may induce up to ℓ_r "signal" eigenvalues separated from the eigenvalue distribution of $\mathring{\Sigma}_r$. Our results should be interpreted in the setting where $\mathring{\Sigma}_r$ itself asymptotically does not have additional outlier eigenvalues which separate from the bulk of its eigenvalue distribution.

As $\alpha_1, \dots, \alpha_k$ are unobserved in this model, one cannot construct a direct sample covariance estimator for $\Sigma_1, \dots, \Sigma_k$. Instead, each Σ_r is classically estimated by a MANOVA estimator of the form

$$\widehat{\Sigma} = Y^{\mathsf{T}} B Y, \tag{2.3}$$

where the symmetric matrix B is chosen to satisfy the properties

$$BX = 0, \qquad \mathbb{E}[Y^{\mathsf{T}} B Y] = \Sigma_r.$$

Such an estimator is unbiased and equivariant to rotations of coordinates in $\mathbb{R}^p$. Here, B and $\widehat{\Sigma}$ depend implicitly on $r \in \{1, \dots, k\}$, but we suppress this dependence in the notation.

As in the earlier works [FJ16, FJS18], we study an asymptotic regime summarized as follows.

Assumption 2.2. The dimensions $n, p, n_1, \dots, n_k \to \infty$ where k is a fixed constant. There are universal constants $C, c > 0$ such that, for each $r = 1, \dots, k$,

- $c < p/n < C$ and $c < n_r/n < C$,
- $\|U_r\| < C$ and $\|B\| < C/n$
- $\|\Gamma_r\| < C$, $\|\mathring{\Sigma}_r\| < C$, and $\ell_r < C$.

Thus, the number of realizations of each random effect is proportional to the dimension p. The scaling $\|B\| < C/n$ is usual for MANOVA estimators, to yield $\widehat{\Sigma}$ on the same scale as Σ_r.

2.2. **Deterministic equivalent measure.** Consider first the setting of no spikes, meaning $\ell_r = 0$ and $\Sigma_r = \mathring{\Sigma}_r$ for each $r = 1, \dots, k$. We introduce the notations

$$F_{rs} = \sqrt{n_r n_s} U_r^{\mathsf{T}} B U_s \in \mathbb{R}^{n_r \times n_s}, \quad F = (F_{rs})_{r,s=1}^k \in \mathbb{R}^{n_+ \times n_+}, \quad n_+ = n_1 + \ldots + n_k, \tag{2.4}$$

$$\operatorname{diag}_n(a) = \operatorname{diag}(a_1 \operatorname{Id}_{n_1}, \dots, a_k \operatorname{Id}_{n_k}) \in \mathbb{R}^{n_+ \times n_+}, \quad b \cdot \mathring{\Sigma} = b_1 \mathring{\Sigma}_1 + \ldots + b_k \mathring{\Sigma}_k \in \mathbb{R}^{p \times p},$$

and Tr_r is the trace of the (r, r) block (of size $n_r \times n_r$) in the $k \times k$ matrix block decomposition corresponding to $\mathbb{C}^{n_+} = \mathbb{C}^{n_1} \oplus \cdots \oplus \mathbb{C}^{n_k}$. We define the Stieltjes transform of a measure μ as $m(z) = \int (x - z)^{-1} d\mu(x)$. The following deterministic approximation for the spectral distribution of $\widehat{\Sigma}$ was established in [FJ16].

Theorem 2.3 ([FJ16])**.** *Suppose Assumptions 2.1 and 2.2 hold, and $\ell_r = 0$ for each $r = 1, \dots, k$. Let $\widehat{\Sigma}$ be as in (2.3), and let $\widehat{\mu} = p^{-1} \sum_{i=1}^p \delta_{\lambda_i(\widehat{\Sigma})}$ be the empirical distribution of its eigenvalues.*

For each $z \in \mathbb{C}^+$, there exist unique z-dependent values $a_1, \ldots, a_k \in \mathbb{C}^+ \cup \{0\}$ and $b_1, \ldots, b_k \in \overline{\mathbb{C}^+}$ that satisfy the equations

$$a_r = -n_r^{-1} \operatorname{Tr}\Big((z \operatorname{Id} + b \cdot \mathring{\Sigma})^{-1} \mathring{\Sigma}_r\Big), \tag{2.5}$$

$$b_r = -n_r^{-1} \operatorname{Tr}_r\Big((\operatorname{Id} + F \operatorname{diag}_n(a))^{-1} F\Big). \tag{2.6}$$

The function $m_0 : \mathbb{C}^+ \to \mathbb{C}^+$ defined by

$$m_0(z) = -p^{-1} \operatorname{Tr}\Big((z \operatorname{Id} + b \cdot \mathring{\Sigma})^{-1}\Big) \tag{2.7}$$

is the Stieltjes transform of a probability measure μ_0 on $\mathbb{R}$, and $\widehat{\mu} - \mu_0 \to 0$ weakly a.s.

The distribution μ_0 is an n-dependent deterministic equivalent measure, defined by the population covariances $\mathring{\Sigma}_1, \ldots, \mathring{\Sigma}_k$ and the structure of the model (but not depending on the realizations of the random effects $\alpha_1, \ldots, \alpha_k$).

We call $(\mathcal{A}, \tau)$ a *von Neumann probability space* (W^*-probability space) if $\mathcal{A}$ is a (n-dependent) von Neumann algebra and $\tau : \mathcal{A} \to \mathbb{C}$ a positive, faithful, normal trace. The proof in [FJ16, Section 4] illustrates that μ_0 is the τ-distribution of an operator w in such a space $(\mathcal{A}, \tau)$, arising in the following way: Set $N = (k+1)p + n_1 + \ldots + n_k$. In (2.1), write

$$E_r = \sqrt{n_r} G_r H_r, \tag{2.8}$$

where $G_r \in \mathbb{R}^{n_r \times p}$ has i.i.d. $\mathcal{N}(0, 1/n_r)$ entries and $H_r = \mathring{\Sigma}^{1/2} \in \mathbb{R}^{p \times p}$. Then, when $\ell_r = 0$ for all r, we obtain

$$W = \sum_{r=1}^{k} \sum_{s=1}^{k} H_r^* G_r^* F_{rs} G_s H_s, \tag{2.9}$$

where $\{F_{rs} : r, s = 1, \ldots, k\}$ are defined in (2.4). Embed $\{F_{rs}, G_r, H_r : r, s = 1, \ldots, k\}$ into $\mathbb{C}^{N \times N}$ by zero-padding, in the following blocks of the $(2k+1) \times (2k+1)$ block decomposition for $\mathbb{C}^N = \mathbb{C}^p \oplus \cdots \oplus \mathbb{C}^p \oplus \mathbb{C}^{n_1} \ldots \oplus \mathbb{C}^{n_k}$:

$$\begin{pmatrix} & H_1^* & \cdots & H_k^* & & & \\ H_1 & & & & G_1^* & & \\ \vdots & & & & & \ddots & \\ H_k & & & & & & G_k^* \\ & G_1 & & & F_{11} & \cdots & F_{1k} \\ & & \ddots & & \vdots & \ddots & \vdots \\ & & & G_k & F_{k1} & \cdots & F_{kk} \end{pmatrix} \tag{2.10}$$

Denote by $\tilde{F}_{rs}, \tilde{H}_r \in \mathbb{C}^{N \times N}$ the zero-padded matrices. Define also the mutually orthogonal projections

$$P_0 = \operatorname{diag}(\operatorname{Id}_p, 0, \ldots, 0), \quad P_1 = (0, \operatorname{Id}_p, \ldots, 0), \quad \ldots, \quad P_{2k} = \operatorname{diag}(0, \ldots, 0, \operatorname{Id}_{n_k}) \in \mathbb{C}^{N \times N}.$$

Let $(\mathcal{A}, \tau)$ be any von Neumann probability space containing mutually orthogonal projections $p_0, p_1, \ldots, p_{2k}$ which satisfy $\tau(p_0) = \ldots = \tau(p_k) = p/N$ and $\tau(p_{k+r}) = n_r/N$ for each $r = 1, \ldots, k$. This is a *rectangular probability space* as defined in [BG09]. Let $(\mathcal{A}, \tau)$ also contain $\{f_{rs}, g_r, h_r : r, s = 1, \ldots, k\}$ such that

(1) $p_{k+r} f_{rs} p_{k+s}$, $p_{k+r} g_r p_r = g_r$, and $p_r h_r p_0 = h_r$.
(2) For any non-commutative $*$-polynomial Q of k variables,
$$\tau(Q(h_1, \ldots, h_k)) = N^{-1} \operatorname{Tr} Q(\tilde{H}_1, \ldots, \tilde{H}_k).$$
Similarly, for any non-commutative $*$-polynomial Q of k^2 variables,
$$\tau(Q(f_{11}, f_{12}, \ldots, f_{kk})) = N^{-1} \operatorname{Tr} Q(\tilde{F}_{11}, \tilde{F}_{12}, \ldots, \tilde{F}_{kk}).$$
(3) For each $r \in \{1, \ldots, k\}$ and $l \geq 0$,
$$\frac{N}{p} \tau((g_r^* g_r)^l) = \int x^l \nu_{p/n_r}(dx)$$
where ν is the Marcenko-Pastur law with parameter p/n_r.
(4) The families $\{f_{rs} : r, s = 1, \ldots, k\}$, $\{h_r : r = 1, \ldots, k\}$, and individual elements $g_1, \ldots, g_k$ are free with amalgamation over the subalgebra $\mathcal{D} = \langle p_0, \ldots, p_{2k} \rangle$.

The elements $f_{rs}, g_r, h_r \in \mathcal{A}$ form a *free deterministic equivalent* for our matrix model, in the sense of [SV12] and [FJ16, Definition 3.8].

The element which models $\widehat{\Sigma}$ is then

$$w = \sum_{r=1}^{k} \sum_{s=1}^{k} h_r^* g_r^* f_{rs} g_s h_s. \tag{2.11}$$

Only the $(0,0)$-block of w is non-zero—i.e. w belongs to the compressed algebra $\mathcal{A}^c = \{a \in \mathcal{A} : a = p_0 a p_0\}$, which has unit p_0 and trace $\tau^c(a) = (N/p)\tau(p_0 a p_0)$. The law μ_0 in Theorem 2.3 is the τ^c-distribution of w, meaning for any continuous function $f : \mathbb{R} \to \mathbb{C}$, we have

$$\int f(x) d\mu_0(x) = \tau^c(f(w))$$

where $f(w)$ is defined by the functional calculus on $\mathcal{A}^c$. Since τ is a faithful trace, so is τ^c as a trace on $\mathcal{A}^c$, and we thus have

$$\operatorname{supp}(\mu_0) = \operatorname{spec}(w) \tag{2.12}$$

where $\operatorname{spec}(w)$ is the spectrum of w as an element of $\mathcal{A}^c$. (See [NS06, Propositions 3.13 and 3.15].)

2.3. **First-order behavior of principal components.** Theorem 2.3 establishes the approximation by μ_0 at the level of weak convergence, and in particular does not guarantee the absence of outlier eigenvalues separated from $\operatorname{supp}(\mu_0)$. Denoting by $\operatorname{supp}(\mu_0)_\delta$ the δ-neighborhood

$$\operatorname{supp}(\mu_0)_\delta = \{x \in \mathbb{R} : \operatorname{dist}(x, \operatorname{supp}(\mu_0)) < \delta\},$$

our first main result provides such a guarantee.

Theorem 2.4. *Suppose Assumptions 2.1 and 2.2 hold, and $\ell_r = 0$ for each $r = 1, \dots, k$. Let $\widehat{\Sigma}$ be as in (2.3), with spectrum $\operatorname{spec}(\widehat{\Sigma})$. Then for any constant $\delta > 0$, almost surely for all large n,*

$$\operatorname{spec}(\widehat{\Sigma}) \subset \operatorname{supp}(\mu_0)_\delta.$$

We will prove Theorem 2.4 using a strong asymptotic freeness result for GOE and deterministic matrices (see Section 3) and an embedding argument. We defer all proofs of results in this section to Section 5.

When $\ell_r \neq 0$ for some $r \in \{1, \dots, k\}$, there may be "outlier" eigenvalues of $\widehat{\Sigma}$ that separate from the bulk distribution μ_0 described in the preceding section. We now quantify the first-order limiting behavior of these outlier eigenvalues and the alignments of the associated eigenvectors with the true signal directions. Our description will be in terms of the quantities $\{a_r\}_{r=1}^k$ and $\{b_r\}_{r=1}^k$ from Theorem 2.3, which we extend to $\mathbb{C} \setminus \operatorname{supp}(\mu_0)$ in the following, whose proof is deferred to Section 5.1.

Proposition 2.5. *For any positive semidefinite $\mathring{\Sigma}_1, \dots, \mathring{\Sigma}_k \in \mathbb{R}^{p \times p}$ and symmetric $F \in \mathbb{R}^{M \times M}$, let μ_0 be the measure defined by (2.5–2.7). Then the z-dependent values $a_1, \dots, a_k, b_1, \dots, b_k, m_0$ which solve (2.5–2.7) extend analytically to functions on $\mathbb{C} \setminus \operatorname{supp}(\mu_0)$. The matrices $z \operatorname{Id} + b \cdot \mathring{\Sigma}$ and $\operatorname{Id} + F \operatorname{diag}_n(a)$ are invertible on all of $\mathbb{C} \setminus \operatorname{supp}(\mu_0)$, and these extensions satisfy (2.5–2.7) on all of $\mathbb{C} \setminus \operatorname{supp}(\mu_0)$.*

To state our result on outlier eigenvalues and eigenvectors, introduce the notation

$$\Gamma = \begin{pmatrix} \Gamma_1 \\ \vdots \\ \Gamma_k \end{pmatrix} \in \mathbb{R}^{\ell_+ \times p}, \qquad \ell_+ = \ell_1 + \ldots + \ell_k,$$

$$b \cdot \mathring{\Sigma} = \sum_{r=1}^{k} b_r(\lambda) \mathring{\Sigma}_r, \qquad \operatorname{diag}_\ell(b) = \operatorname{diag}(b_1(\lambda) \operatorname{Id}_{\ell_1}, \dots, b_k(\lambda) \operatorname{Id}_{\ell_k}),$$

and set

$$T(\lambda) = \operatorname{Id} + \Gamma(\lambda \operatorname{Id} + b \cdot \mathring{\Sigma})^{-1} \Gamma^{\mathsf{T}} \operatorname{diag}_\ell(b) \in \mathbb{R}^{\ell_+ \times \ell_+}. \tag{2.13}$$

Define the multiset of roots of T by

$$\Lambda_0 = [\lambda \in \mathbb{R} \setminus \operatorname{supp}(\mu_0) : 0 = \det T(\lambda)] \tag{2.14}$$

counting their analytic multiplicities. For two finite multisets $A, B \subset \mathbb{R}$, define

$$\text{ordered-dist}(A,B) = \begin{cases} \infty & \text{if } |A| \neq |B| \\ \max_i\{|a_{(i)} - b_{(i)}|\} & \text{if } |A| = |B|, \end{cases}$$

where $a_{(i)}$ and $b_{(i)}$ are the ordered values of A and B counting multiplicity.

Theorem 2.6. *Suppose Assumptions 2.1 and 2.2 hold, let $\widehat{\Sigma}$ be as in (2.3), and let Λ_0 be defined by (2.14). Fix any constant $\delta > 0$. Almost surely as $n \to \infty$, there exist $\Lambda_\delta \subseteq \Lambda_0$ and $\widehat{\Lambda}_\delta \subseteq \text{spec}(\widehat{\Sigma})$ containing all elements of Λ_0 and $\text{spec}(\widehat{\Sigma})$ outside $\text{supp}(\mu_0)_\delta$, such that*

$$\text{ordered-dist}(\Lambda_\delta, \widehat{\Lambda}_\delta) \to 0.$$

Theorem 2.7. *In the setting of Theorem 2.6, pick any $\lambda \in \Lambda_\delta$ of multiplicity 1 such that $|\lambda - \lambda'| > \delta$ for all other $\lambda' \in \Lambda_\delta$. Let $u \in \ker T(\lambda)$ be a unit vector, and let $\widehat{v}$ be the unit eigenvector corresponding to the eigenvalue $\widehat{\lambda}$ of $\widehat{\Sigma}$ closest to λ. Almost surely as $n \to \infty$, for the appropriate choice of sign of $\widehat{v}$,*

$$\Gamma \widehat{v} - \alpha^{-1/2} u \to 0,$$

where

$$\alpha = u^\mathsf{T}\Big(- \text{diag}_\ell(b(\lambda))\Gamma \cdot \partial_\lambda[(\lambda \,\text{Id} + b(\lambda)\cdot \mathring{\Sigma})^{-1}] \cdot \Gamma^\mathsf{T} \,\text{diag}_\ell(b(\lambda)) + \text{diag}_\ell(b'(\lambda))\Big) u. \tag{2.15}$$

We state Theorem 2.6 as a matching of points in $\text{spec}(\widehat{\Sigma})$ and Λ_0, rather than convergence of $\text{spec}(\widehat{\Sigma})$ to Λ_0, as Λ_0 is a deterministic but also n-dependent set. In the proof of Theorem 2.7, we will establish that the unit vector $u \in \ker T(\lambda)$ is unique up to sign.

Qualitatively, Theorems 2.6 and 2.7 imply similar phenomena as occur in the isotropic noise setting studied in [FJS18]: The outlier eigenvalue locations are predicted by the roots of a determinant equation, which depends on all of the noise covariances $\mathring{\Sigma}_1, \dots, \mathring{\Sigma}_k$ via the analytic functions $b_1(\lambda), \dots, b_k(\lambda)$ defining μ_0, as well as on the alignments between the signal directions and these noise covariances via $\Gamma(\lambda + b\cdot\mathring{\Sigma})^{-1}\Gamma^\mathsf{T}$. An aliasing effect may occur, in which large signal eigenvalues Γ_s in one covariance Σ_s may yield outliers in the estimate of a different covariance Σ_r. The number of outliers is predicted by the cardinality $|\Lambda_0|$, which implicitly describes a phase transition phenomenon. If $\|\Gamma_1\|, \dots, \|\Gamma_k\|$ are all small enough, then $0 = \det T(\lambda)$ has no roots, and $\widehat{\Sigma}$ will (with high probability) have no outliers. The thresholds at which outliers appear depend not only on the signal sizes in each individual $\Gamma_1, \dots, \Gamma_k$, but on the alignments of the signal vectors across different variance components and with the noise covariances. Theorem 2.7 characterizes the asymptotic alignment between the estimated principal component $\widehat{v}$ and each true signal direction (row) in Γ. As shown in [FJS18], the estimate $\widehat{v}$ in covariance Σ_r may be biased towards signal directions of a different covariance Σ_s.

Remark 2.8. As expected, the biases exhibited in Theorems 2.6 and 2.7 vanish for MANOVA estimators in the limit of large signal size. Consider, for simplicity, a limit where the first row v of Γ has $\|v\| \to \infty$, while the remaining rows of Γ are fixed. Thus Σ_1 has a signal eigenvalue $\lambda_0 \sim \|v\|^2$ as $\|v\| \to \infty$. For $\lambda \to \infty$, we may verify from (2.5–2.6) that for all s, we have

$$a_s(\lambda) \to 0, \qquad b_s(\lambda) \to -n_s^{-1} \,\text{Tr}_s\, F.$$

If $\widehat{\Sigma} = Y^\mathsf{T} B Y$ is a MANOVA estimator of Σ_1, for which $\mathbb{E}[Y^\mathsf{T} B Y] = \Sigma_1$, then this implies $U_1^\mathsf{T} B U_1 = \text{Id}$ and $U_s^\mathsf{T} B U_s = 0$ for all $s \neq 1$. Thus, recalling the definition of F in (2.4),

$$b_1(\lambda) \to -1, \qquad b_s(\lambda) \to 0 \text{ for all } s \neq 1.$$

Then $(\lambda + b\cdot\mathring{\Sigma})^{-1} \sim \lambda^{-1}\,\text{Id}$, and $T(\lambda) \sim \text{Id} - \lambda_0 e_1 e_1^\mathsf{T}/\lambda$ where $e_1 \in \mathbb{R}^{\ell_+}$ is the first standard basis vector. So $0 = \det T(\lambda)$ has a root $\lambda \sim \lambda_0$ as $\|v\| \to \infty$, and Theorem 2.6 implies $\widehat{\Sigma}$ has an outlier eigenvalue $\widehat{\lambda} \sim \lambda_0$.

The vector $u \in \ker T(\lambda)$ approaches e_1, so $\Gamma^\mathsf{T}\,\text{diag}_\ell(b)u = -v + o(\|v\|)$ as $\|v\| \to \infty$. We may further verify for all s that

$$|a_s'(\lambda)| \lesssim \lambda^{-2}, \qquad |b_s'(\lambda)| \lesssim \lambda^{-2}.$$

Then $\partial_\lambda[(\lambda\,\text{Id} + b\cdot\mathring{\Sigma})^{-1}] \sim \lambda^{-2}\,\text{Id} \sim \lambda_0^{-2}\,\text{Id}$, so $\alpha \sim \lambda_0^{-2}\|v\|^2 \sim \lambda_0^{-1}$. Theorem 2.7 then gives $|\langle v, \widehat{v}\rangle| \sim \lambda_0^{1/2} \sim \|v\|$, so that $\widehat{v}$ is perfectly aligned with v in the large $\|v\|$ limit.

Remark 2.9. In the setting of isotropic noise, meaning $\mathring{\Sigma}_r = \sigma_r^2 \operatorname{Id}$ for each $r = 1, \dots, k$, Theorems 2.6 and 2.7 coincide with results of [FJS18]. Indeed, comparing (2.7) with (2.5), we have in this setting $a_r(z) = (p\sigma_r^2/n_r)m_0(z)$ for each r. Then $b_r(z)$ coincides with $-t_r(z)$ as defined in [FJS18, Eq. (3.2)]. (Note that the matrix F_{rs} in [FJS18] corresponds to $(p\sigma_r\sigma_s/\sqrt{n_r n_s})F_{rs}$ in the notation of this paper.) Applying $\det(\operatorname{Id} + XY) = \det(\operatorname{Id} + YX)$, our determinant equation $0 = \det T(\lambda)$ is equivalent to

$$0 = \det\Big(\operatorname{Id} + (\lambda \operatorname{Id} + b \cdot \mathring{\Sigma})^{-1}\Gamma^{\mathsf{T}} \operatorname{diag}_\ell(b)\Gamma\Big) \Leftrightarrow 0 = \det(\lambda \operatorname{Id} + b \cdot \mathring{\Sigma} + \Gamma^{\mathsf{T}} \operatorname{diag}_\ell(b)\Gamma) = \det(\lambda \operatorname{Id} + b \cdot \Sigma).$$

This is the same as the equation defining Λ_0 in [FJS18, Eq. (3.4)].

For the eigenvectors, note that $u \in \ker T(\lambda)$ corresponds to

$$v = -M(\lambda)u \in \ker(\operatorname{Id} + b \cdot \Sigma), \quad M(\lambda) = (\lambda \operatorname{Id} + b \cdot \mathring{\Sigma})^{-1}\Gamma^{\mathsf{T}} \operatorname{diag}_\ell(b).$$

Then, in our notation, [FJS18, Theorem 3.3] implies

$$\Gamma \widehat{v} + \Big(u^{\mathsf{T}} M(\lambda)^{\mathsf{T}} \Big(\operatorname{Id} + b'(\lambda) \cdot \mathring{\Sigma} + \Gamma^{\mathsf{T}} \operatorname{diag}_\ell(b')\Gamma\Big) M(\lambda) u\Big)^{-1/2} \Gamma M(\lambda) u \to 0.$$

Since $u \in \ker T(\lambda)$, we have $\Gamma M(\lambda)u = -u$. Applying this and simplifying, we recover exactly Theorem 2.7.

3. Strong asymptotic freeness of GOE and deterministic matrices

Fix integers $p, q \geq 0$. Let $X_1, \dots, X_p \in \mathbb{R}^{N \times N}$ be independent GOE matrices, with diagonal entries distributed as $\mathcal{N}(0, 2/N)$ and off-diagonal entries as $\mathcal{N}(0, 1/N)$. Let $Y_1, \dots, Y_q \in \mathbb{C}^{N \times N}$ be deterministic matrices. Write as shorthand $M_N = \mathbb{C}^{N \times N}$, and denote by $\operatorname{tr}_N = N^{-1} \operatorname{Tr}$ the normalized matrix trace on M_N. Denote $\mathbf{X}_N = (X_1, \dots, X_p)$ and $\mathbf{Y}_N = (Y_1, \dots, Y_q)$.

Consider an N-dependent von Neumann algebra $\mathcal{A}_N$ with a positive, faithful, normal trace τ_N. Suppose $\mathcal{A}_N$ contains $x_1, \dots, x_p$ and $Y_1, \dots, Y_q$ such that $x_1, \dots, x_p$ are free semicircular elements under τ_N, also free of $Y_1, \dots, Y_q$, and $\tau_N \equiv \operatorname{tr}_N$ restricted to the von Neumann subalgebra $\langle Y_1, \dots, Y_q \rangle$. Denote $\mathbf{x} = (x_1, \dots, x_p)$. The following is the main result of this section.

Theorem 3.1. *Suppose that there exists a constant $C > 0$ such that $\|Y_j\| \leq C$ for all $j = 1, \dots, q$ and all N. Then for any fixed non-commutative self-adjoint $*$-polynomial Q in $p + q$ variables and any constant $\delta > 0$, almost surely for all large N,*

$$\operatorname{spec}(Q(\mathbf{X}_N, \mathbf{Y}_N)) \subset \operatorname{spec}(Q(\mathbf{x}, \mathbf{Y}_N))_\delta. \tag{3.1}$$

Here, $\operatorname{spec}(Q(\mathbf{x}, \mathbf{Y}_N))_\delta$ is the δ-neighborhood of the spectrum of the operator $Q(\mathbf{x}, \mathbf{Y}_N) \in \mathcal{A}_N$.

For our application, we will apply strong asymptotic freeness directly in the above form. However, we observe that we may also obtain the following corollary by the arguments of [Mal12, Section 7].

Theorem 3.2. *Let $\mathbf{x} = (x_1, \dots, x_p)$ and $\mathbf{y} = (y_1, \dots, y_q)$ be elements of a fixed von Neumann algebra $\mathcal{A}$ with a positive, faithful, normal trace τ, such that $\mathbf{x}$ is a free semicircular family free from $\mathbf{y}$. Assume that almost surely as $N \to \infty$, for any $*$-polynomial P in q variables,*

$$\operatorname{tr}_N[P(\mathbf{Y}_N)] \to \tau(P(\mathbf{y})) \quad \text{and} \quad \|P(\mathbf{Y}_N)\| \to \|P(\mathbf{y})\|.$$

Then, almost surely for any $*$-polynomial Q in $p + q$ variables,

$$\operatorname{tr}_N[Q(\mathbf{X}_N, \mathbf{Y}_N)] \to \tau(Q(\mathbf{x}, \mathbf{y})) \quad \text{and} \quad \|Q(\mathbf{X}_N, \mathbf{Y}_N)\| \to \|Q(\mathbf{x}, \mathbf{y})\|. \tag{3.2}$$

Note that upon replacing Y_j and Y_j^* by $(Y_j + Y_j^*)/2$ and $(Y_j - Y_j^*)/(2i)$, we can and will assume without loss of generality that $Y_1, \dots, Y_q$ are Hermitian.

The proof follows an argument analogous to [HT05, Mal12], which established such a result for GUE and GUE + deterministic matrices, respectively. Several modifications to the argument are needed, which are inspired by [Sch05, BC17] that established this result for GOE and complex Wigner + deterministic matrices, respectively. We note that the result of [BC17] requires real and imaginary parts of the complex Wigner matrices to have the same variance, and does not directly apply to the GOE setting.

We provide here a brief outline of the proof and its relation to these previous works:

(1) By the linearization trick of [HT05, Section 2], we first study linear polynomials L with $k \times k$ Hermitian matrix-valued coefficients, for arbitrary fixed dimension k. We aim to show the spectral inclusion (3.1) for such L, see Lemma 3.15.

(2) For this, it suffices to show that the difference between the Cauchy transform of $L(\mathbf{X}_N, \mathbf{Y}_N)$ and of a deterministic measure $\mu_{\mathcal{A}}$ with the same spectrum as $L(\mathbf{x}, \mathbf{Y}_N)$ is at most $\mathrm{poly}((\mathrm{Im}\,\lambda)^{-1})/N^{1+\kappa}$, for some $\kappa > 0$ and any spectral argument $\lambda \in \mathbb{C}^+$. For simplicity, we drop the λ-dependence here and denote this as $O(1/N^{1+\kappa})$. As in [HT05, Mal12], we bound the expected difference by $O(1/N^2)$ and the variance by $O(1/N^4)$; the latter uses the same Gaussian Poincaré argument as in these works.

(3) To bound the expected difference, we work with the expected M_k-valued Cauchy transform $G_{S_N+T_N}$ of $L(\mathbf{X}_N, \mathbf{Y}_N)$, and the M_k-valued Cauchy transform G_{s+T_N} of $L(\mathbf{x}, \mathbf{Y}_N)$. The latter satisfies the operator-valued subordination equation for the free additive convolution,

$$G_{s+T_N}(\Lambda) = G_{T_N}(\Lambda - \mathcal{R}_s(G_{s+T_N}(\Lambda))).$$

Applying a similar Gaussian integration-by-parts argument as in [Mal12], we show

$$G_{S_N+T_N}(\Lambda) = G_{T_N}(\Lambda - \mathcal{R}_s(G_{S_N+T_N}(\Lambda))) + O(1/N),$$

see Lemma 3.5. In contrast to the GUE setting of [Mal12], this is a *first-order* remainder of size $O(1/N)$, not $O(1/N^2)$ as required. The $O(1/N)$ term vanishes for the GUE by a cancellation due to the real and imaginary parts having the same variance, but does not vanish for the GOE. A similar difficulty occurred also in [Sch05].

(4) The bulk of the additional work in our argument lies in obtaining the second-order $O(1/N^2)$ approximation. In Proposition 3.11 below, applying the stability property of the subordination equation established in [Mal12, Proposition 4.3] together with a Taylor expansion of G_{T_N}, we obtain

$$\|G_{S_N+T_N}(\Lambda) - G_{s+T_N}(\Lambda) - \mathcal{L}_\Lambda(R_N(\Lambda))\| \leq O(1/N^2),$$

where $\|\mathcal{L}_\Lambda(R_N(\Lambda))\| \leq O(1/N)$. We approximate the random quantity $R_N(\Lambda)$ by a deterministic approximation $R_{\mathcal{A}}(\Lambda)$, and show that $G_{s+T_N}(\Lambda) + \mathcal{L}_\Lambda(R_{\mathcal{A}}(\Lambda))$ is the Cauchy transform of a deterministic measure $\mu_{\mathcal{A}}$ as above. For the approximation $R_N(\Lambda) \approx R_{\mathcal{A}}(\Lambda)$, we follow an approach inspired by [Sch05], and we identify the key term of $R_N(\Lambda) - R_{\mathcal{A}}(\Lambda)$ as the derivative of the difference of certain "left-augmented" M_{2k}-valued Cauchy transforms of $L(\mathbf{X}_N, \mathbf{Y}_N)$ and $L(\mathbf{x}, \mathbf{Y}_N)$ in an expanded $2k \times 2k$ coefficient space, see (3.5) below. We bound this difference using a left-augmented subordination identity for $R_{\mathcal{A}}(\Lambda)$, an approximate such identity for $R_N(\Lambda)$, and a second application of the stability property of [Mal12, Proposition 4.3].

We remark that a term similar to $\mathcal{L}_\Lambda(R_N(\Lambda))$ above appeared also in [BC17, Theorem 5.7], but arose from the remainders corresponding to the third-and-higher-order cumulants in the integration-by-parts identity applied to non-Gaussian variables, rather than from the difference in variance between the real and imaginary parts.

(5) Finally, having established (3.1) for all such linear polynomials L, we may directly establish (3.1) for all Q by applying the linearization and ultraproduct argument of [HT05, Section 7] in a subsequential form. This concludes the proof of Theorem 3.1.

In the remainder of this section, we carry out these steps. The implication Theorem 3.2 is deferred to the end of the section.

3.1. **Augmented Cauchy and R-transforms.** In this section, we review some constructions from free probability theory and introduce the left-augmented transforms described above. Let $(\mathcal{A}, \tau)$ be any von Neumann probability space. For a von Neumann subalgebra $\mathcal{B} \subset \mathcal{A}$, denote by

$$\tau^{\mathcal{B}} : \mathcal{A} \to \mathcal{B}$$

the (unique) conditional expectation satisfying the τ-invariance $\tau(\tau^{\mathcal{B}}(a)) = \tau(a)$. For each $l \geq 1$, let $\mathrm{NC}(l)$ be the space of non-crossing partitions of $1, \ldots, l$. For $\pi \in \mathrm{NC}(l)$, denote by $\kappa_\pi^{\mathcal{B}}(a_1, \ldots, a_l)$ the non-crossing cumulant corresponding to π. These satisfy the moment-cumulant relations

$$\tau^{\mathcal{B}}(a_1 \ldots a_l) = \sum_{\pi \in \mathrm{NC}(l)} \kappa_\pi^{\mathcal{B}}(a_1, \ldots, a_l).$$

Denote by

$$G_a^{\mathcal{B}}(b) = \tau^{\mathcal{B}}((b-a)^{-1}) = \sum_{l \geq 0} b^{-1}(ab^{-1})^l, \qquad \mathcal{R}_a^{\mathcal{B}}(b) = \sum_{l \geq 1} \kappa_l^{\mathcal{B}}(a, ba, \ldots, ba)$$

the $\mathcal{B}$-valued Cauchy- and $\mathcal{R}$-transforms of $a \in \mathcal{A}$, the former defined for all invertible $b \in \mathcal{B}$ with $\|b^{-1}\|$ sufficiently small, and the latter for all $b \in \mathcal{B}$ with $\|b\|$ sufficiently small. (Note that, following conventions in free probability, we take the opposite sign for $G_a^{\mathcal{B}}(b)$ here as for the Stieltjes transform defined in Section 2.1.) The moment-cumulant relations yield the identity

$$G_a^{\mathcal{B}}(b) = (b - \mathcal{R}_a^{\mathcal{B}}(G_a^{\mathcal{B}}(b)))^{-1} \tag{3.3}$$

for invertible $b \in \mathcal{B}$ with $\|b^{-1}\|$ sufficiently small. If $s, t \in \mathcal{A}$ are free with amalgamation over $\mathcal{B}$, this yields also the subordination identity for $G_{s+t}^{\mathcal{B}}$, given by

$$G_{s+t}^{\mathcal{B}}(b) = G_t^{\mathcal{B}}(b - \mathcal{R}_s^{\mathcal{B}}(G_{s+t}^{\mathcal{B}}(b))). \tag{3.4}$$

See [MS17, Chapter 9] for additional details.

In this section, as well as in Section 5, we will make use of the following "left-augmented" Cauchy- and R-transforms, defined for $a_1, a \in \mathcal{A}$ and $b \in \mathcal{B}$ by the mixed moments and mixed cumulants

$$G_{a_1,a}^{\mathcal{B}}(b) = \tau^{\mathcal{B}}\Big(a_1(b-a)^{-1}\Big) = \sum_{l \geq 0} \tau^{\mathcal{B}}(a_1 b^{-1}(ab^{-1})^l), \tag{3.5}$$

$$R_{a_1,a}^{\mathcal{B}}(b) = \sum_{l \geq 1} \kappa_l^{\mathcal{B}}(a_1, ba, \ldots, ba). \tag{3.6}$$

The following result, analogous to (3.3), is also a consequence of the moment-cumulant relations.

Lemma 3.3. *For $a_1, a \in \mathcal{A}$ and all invertible $b \in \mathcal{B}$ with $\|b^{-1}\|$ sufficiently small,*

$$G_{a_1,a}^{\mathcal{B}}(b) = R_{a_1,a}^{\mathcal{B}}(G_a^{\mathcal{B}}(b)) G_a^{\mathcal{B}}(b). \tag{3.7}$$

Proof. We apply the cumulant expansion to obtain

$$G_{a_1,a}^{\mathcal{B}}(b) = \sum_{l \geq 0} \tau^{\mathcal{B}}(a_1 b^{-1}(ab^{-1})^l) = \sum_{l \geq 0} \sum_{\pi \in \mathrm{NC}(l+1)} \kappa_\pi^{\mathcal{B}}(a_1 b^{-1}, ab^{-1}, \ldots, ab^{-1}). \tag{3.8}$$

For a given non-crossing partition $\pi \in \mathrm{NC}(l+1)$, let $S \in \pi$ denote the element containing 1. Then the size m of S can range from 1 to $l+1$. Denote $S = \{j_0, j_1, \ldots, j_{m-1}\}$ where $j_0 = 1$. Set $c_i = j_i - j_{i-1} - 1$ for $i = 1, \ldots, m-1$ to be the number of elements between j_{i-1} and j_i, and set $c_m = l + 1 - j_{m-1}$ as the number of elements after j_{m-1}. Then $c_1, \ldots, c_m$ sum to $l+1-m$, and the remaining elements of π form non-crossing partitions of these intervals of sizes $c_1, \ldots, c_m$. Hence, applying the definition and multilinearity of $\kappa_\pi^{\mathcal{B}}$, we have

$$\begin{aligned}
&\sum_{\pi \in \mathrm{NC}(l+1)} \kappa_\pi^{\mathcal{B}}(a_1, \ldots, a_{l+1}) \\
&= \sum_{m=1}^{l+1} \sum_{\substack{c_1,\ldots,c_m \geq 0 \\ \sum_i c_i = l+1-m}} \sum_{\pi_1 \in \mathrm{NC}(c_1), \ldots, \pi_m \in \mathrm{NC}(c_m)} \\
&\qquad \kappa_m^{\mathcal{B}}(a_1 \kappa_{\pi_1}^{\mathcal{B}}(a_2, \ldots, a_{j_1-1}), a_{j_1} \kappa_{\pi_2}^{\mathcal{B}}(a_{j_1+1}, \ldots, a_{j_2-1}), \ldots, a_{j_{m-1}} \kappa_{\pi_m}^{\mathcal{B}}(a_{j_{m-1}+1}, \ldots, a_{l+1})) \\
&= \sum_{m=1}^{l+1} \sum_{\substack{c_1,\ldots,c_m \geq 0 \\ \sum_i c_i = l+1-m}} \kappa_m^{\mathcal{B}}(a_1 \tau^{\mathcal{B}}(a_2 \ldots a_{j_1-1}), a_{j_1} \tau^{\mathcal{B}}(a_{j_1+1} \ldots a_{j_2-1}), \ldots, a_{j_{m-1}} \tau^{\mathcal{B}}(a_{j_{m-1}+1} \ldots a_{l+1})).
\end{aligned}$$

Applying this to (3.8), exchanging orders of summations by

$$\sum_{l \geq 0} \sum_{m=1}^{l+1} \sum_{\substack{c_1,\ldots,c_m \geq 0 \\ \sum_i c_i = l+1-m}} = \sum_{m \geq 1} \sum_{l \geq m-1} \sum_{\substack{c_1,\ldots,c_m \geq 0 \\ \sum_i c_i = l+1-m}} = \sum_{m \geq 1} \sum_{c_1 \geq 0, \ldots, c_m \geq 0},$$

and then applying the definition of $\mathcal{B}$-valued Cauchy and R-transforms, we obtain

$$\begin{aligned} G^{\mathcal{B}}_{a_1,a}(b) &= \sum_{l\geq 0}\sum_{m=1}^{l+1} \sum_{\substack{c_1,\ldots,c_m\geq 0 \\ \sum_i c_i = l+1-m}} \kappa^{\mathcal{B}}_m(a_1 b^{-1}\tau^{\mathcal{B}}((ab^{-1})^{c_1}), ab^{-1}\tau^{\mathcal{B}}((ab^{-1})^{c_2}), \ldots, ab^{-1}\tau^{\mathcal{B}}((ab^{-1})^{c_m})) \\ &= \sum_{m\geq 1} \kappa^{\mathcal{B}}_m(a_1 G^{\mathcal{B}}_a(b), aG^{\mathcal{B}}_a(b), \ldots, a)G^{\mathcal{B}}_a(b) \\ &= R^{\mathcal{B}}_{a_1,a}(G^{\mathcal{B}}_a(b))G^{\mathcal{B}}_a(b). \end{aligned}$$

For $\|b^{-1}\|$ sufficiently small, the preceding infinite series are all absolutely norm-convergent, and hence the preceding manipulations are valid as convergent series in $\mathcal{B}$. □

This leads also to the following subordination identity for the above left-augmented transforms.

Lemma 3.4 (Left subordination identity)**.** *Suppose $s, t, m \in \mathcal{A}$ are such that s and $\{t, m\}$ are free with amalgamation over $\mathcal{B}$. Then for any invertible $b \in \mathcal{B}$ with $\|b^{-1}\|$ sufficiently small,*

$$G^{\mathcal{B}}_{m,s+t}(b) = G^{\mathcal{B}}_{m,t}(b - \mathcal{R}^{\mathcal{B}}_s(G^{\mathcal{B}}_{s+t}(b))).$$

Proof. Denote $b' = G^{\mathcal{B}}_{s+t}(b) \in \mathcal{B}$. The usual subordination identity gives $b' = G^{\mathcal{B}}_t(b - \mathcal{R}^{\mathcal{B}}_s(b'))$. Then

$$\begin{aligned} G^{\mathcal{B}}_{m,s+t}(b) &= \mathcal{R}^{\mathcal{B}}_{m,s+t}(b')b' \\ &= \sum_{l\geq 1} \kappa^{\mathcal{B}}_l(m, b'(s+t), \ldots, b'(s+t))b' \\ &= \sum_{l\geq 1} \kappa^{\mathcal{B}}_l(m, b't, \ldots, b't)b' \\ &= \mathcal{R}^{\mathcal{B}}_{m,t}(b')b' \\ &= G^{\mathcal{B}}_{m,t}(b - \mathcal{R}^{\mathcal{B}}_s(b')) \end{aligned}$$

where the first and last equalities apply (3.7) with $a = s + t$ and $a = t$, the second and fourth equalities apply the definition of $\mathcal{R}^{\mathcal{B}}_{a_1,a}$, and the middle equality applies multi-linearity of κ_l, $\mathcal{B}$-freeness of s and m, and vanishing of mixed cumulants for free elements. □

3.2. **Linearization and first-order approximation.** We first consider linear polynomials with matrix-valued coefficients. Fix any $k \geq 1$ and Hermitian matrices $a_0, ..., a_p, b_1, ..., b_q \in M_k$. Set

$$L_N = a_0 \otimes \mathrm{Id}_N + S_N + T_N, \qquad S_N = \sum_{j=1}^{p} a_j \otimes X_j, \qquad T_N = \sum_{j=1}^{q} b_j \otimes Y_j. \tag{3.9}$$

Define correspondingly

$$L_{\mathcal{A}} = a_0 \otimes \mathrm{Id}_N + s + T_N, \qquad s = \sum_{j=1}^{p} a_j \otimes x_j. \tag{3.10}$$

These belong to von Neumann probability spaces $(M_k \otimes M_N, \mathrm{tr}_k \otimes \mathrm{tr}_N)$ and $(M_k \otimes \mathcal{A}_N, \mathrm{tr}_k \otimes \tau_N)$. We denote by Id_k the identity in M_k, and by Id_N both the identity in M_N and the unit in $\mathcal{A}_N$. The space M_k is identified as a subalgebra of both $M_k \otimes M_N$ and $M_k \otimes \mathcal{A}_N$ via the inclusion map $x \mapsto x \otimes \mathrm{Id}_N$, with the partial traces $\mathrm{Id}_k \otimes \mathrm{tr}_N$ and $\mathrm{Id}_k \otimes \tau_N$ being the conditional expectations onto this subalgebra. Throughout, we let $C, C_1, C_2, \ldots$ be arbitrary constants depending on $k, p, q, \|a_0\|, \ldots, \|a_p\|, \|b_1\|, \ldots, \|b_q\|$.

For any element x of a von Neumann algebra $\mathcal{A}$, define the self-adjoint element

$$\mathrm{Im}\, x = \frac{x - x^*}{2i} \in \mathcal{A}.$$

We will use repeatedly the fact that for any self-adjoint element $y \in \mathcal{A}$,

$$\|(x+y)^{-1}\| \leq \|(\mathrm{Im}(x+y))^{-1}\| = \|(\mathrm{Im}\, x)^{-1}\|,$$

see [HT05, Lemma 3.1]. Let

$$M_k^+ = \{X \in M_k : \mathrm{Im}\, X \succ 0\}, \quad M_k^- = \{X \in M_k : \mathrm{Im}\, X \prec 0\}$$

where $\succ$ and $\prec$ denote the positive-definite partial ordering for Hermitian matrices.

For $\Lambda, \Gamma \in M_k^+$, define the resolvents

$$h_{S_N+T_N}(\Lambda) = (\Lambda \otimes \mathrm{Id}_N - S_N - T_N)^{-1}, \quad g_{S_N+T_N}(\Lambda) = \mathbb{E}[h_{S_N+T_N}(\Lambda)], \quad g_{T_N}(\Gamma) = (\Gamma \otimes \mathrm{Id}_N - T_N)^{-1}.$$

Define the M_k-valued Cauchy transforms

$$H_{S_N+T_N}(\Lambda) = (\mathrm{Id}_k \otimes \mathrm{tr}_N)[h_{S_N+T_N}(\Lambda)], \quad G_{S_N+T_N}(\Lambda) = \mathbb{E}[H_{S_N+T_N}(\Lambda)], \quad G_{T_N}(\Gamma) = (\mathrm{Id}_k \otimes \mathrm{tr}_N)[g_{T_N}(\Gamma)].$$

We will eventually apply these with $\Lambda = \lambda \mathrm{Id}_k - a_0$ to obtain the scalar-valued Cauchy transform of L_N. Since S_N and T_N are Hermitian, we have the operator-norm bounds

$$\|H_{S_N+T_N}(\Lambda)\| \le \|h_{S_N+T_N}(\Lambda)\| \le \|(\mathrm{Im}\,\Lambda)^{-1}\|,$$

and similarly for G_{T_N} and g_{T_N}. One may verify that $H_{S_N+T_N}$, $G_{S_N+T_N}$, and G_{T_N} are analytic maps from M_k^+ to M_k^-.

Correspondingly, define the resolvent and M_k-valued Stieltjes transform of $s + T_N$, for $\Lambda \in M_k^+$, by

$$g_{s+T_N}(\Lambda) = (\Lambda \otimes \mathrm{Id}_N - s - T_N)^{-1}, \quad G_{s+T_N}(\Lambda) = (\mathrm{Id}_k \otimes \tau_N)[g_{s+T_N}(\Lambda)].$$

Then by (3.4), G_{s+T_N} satisfies the subordination identity

$$G_{s+T_N}(\Lambda) = G_{T_N}(\Lambda - \mathcal{R}_s(G_{s+T_N}(\Lambda))), \tag{3.11}$$

where

$$\mathcal{R}_s(x) = \sum_{j=1}^{p} a_j x a_j \tag{3.12}$$

is the M_k-valued $\mathcal{R}$-transform of s, see [Mal12, Proposition 4.2]. This identity holds for all $\Lambda \in M_k^+$, as both sides are analytic over M_k^+. Since the a_j's are Hermitian, $x \in M_k^+$ implies $\mathrm{Im}\,\mathcal{R}_s(x) \succeq 0$, and $x \in M_k^-$ implies $\mathrm{Im}\,\mathcal{R}_s(x) \preceq 0$.

The subordination property (3.11) arises from freeness of s and T_N over M_k. In this subsection, we establish the following matrix analogue of this identity, which arises from the asymptotic freeness of S_N and T_N.

Lemma 3.5 (Matrix subordination identity)**.** *Fix any $\Lambda \in M_k^+$, and set $\Gamma = \Lambda - \mathcal{R}_s(G_{S_N+T_N}(\Lambda))$. Then*

$$G_{S_N+T_N}(\Lambda) = G_{T_N}(\Gamma) + R_N(\Lambda, \Gamma, \mathrm{Id}_k \otimes \mathrm{Id}_N) + \Theta_N(\Lambda, \Gamma, \mathrm{Id}_k \otimes \mathrm{Id}_N) \tag{3.13}$$

where

$$\|R_N(\Lambda, \Gamma, \mathrm{Id}_k \otimes \mathrm{Id}_N)\| \le \frac{C}{N}\|(\mathrm{Im}\,\Lambda)^{-1}\|^3, \qquad \|\Theta_N(\Lambda, \Gamma, \mathrm{Id}_k \otimes \mathrm{Id}_N)\| \le \frac{C}{N^2}\|(\mathrm{Im}\,\Lambda)^{-1}\|^5.$$

Comparing with (3.11), there is a "first-order" remainder term R_N and "second-order" remainder term Θ_N, whose exact forms are below. We will further approximate R_N in the next subsection.

We show Lemma 3.5 by specializing the following proposition to $M = \mathrm{Id}_k \otimes \mathrm{Id}_N$.

Proposition 3.6. *For any deterministic $\Lambda, \Gamma \in M_k^+$ and $M \in M_k \otimes M_N$, we have*

$$\mathrm{Id}_k \otimes \mathrm{tr}_N \left(\mathbb{E}[M h_{S_N+T_N}(\Lambda)] - M g_{T_N}(\Gamma) \right) = R_N(\Lambda, \Gamma, M) + \Theta_N(\Lambda, \Gamma, M) \tag{3.14}$$

where

$$\Theta_N(\Lambda, \Gamma, M) = \mathbb{E}\left[\mathrm{Id}_k \otimes \mathrm{tr}_N \left[M g_{T_N}(\Gamma) \Big(\big(\mathcal{R}_s(H_{S_N+T_N}(\Lambda)) - \Lambda + \Gamma \big) \otimes \mathrm{Id}_N \Big) h_{S_N+T_N}(\Lambda) \right] \right],$$

$$R_N(\Lambda, \Gamma, M) = -\frac{1}{N} \sum_{j=1}^{p} \sum_{s,l=1}^{k} \mathbb{E}\left[\mathrm{Id}_k \otimes \mathrm{tr}_N \left[M g_{T_N}(\Gamma)(a_j e_{sl}^{(k)} \otimes \mathrm{Id}_N) h^{\mathsf{T}}_{S_N+T_N}(\Lambda)(e_{sl}^{(k)} a_j \otimes \mathrm{Id}_N) h_{S_N+T_N}(\Lambda) \right] \right].$$

Here, $e_{sl}^{(k)} \in M_k$ is the matrix with (s,l) coordinate equal to 1 and remaining coordinates 0, and $h^{\mathsf{T}}_{S_N+T_N}(\Lambda) = (\Lambda^{\mathsf{T}} \otimes \mathrm{Id}_N - S_N^{\mathsf{T}} - T_N^{\mathsf{T}})^{-1}$ is the (non-conjugated) matrix transpose, where

$$S_N^{\mathsf{T}} = \sum_{j=1}^{p} a_j^{\mathsf{T}} \otimes X_j, \qquad T_N^{\mathsf{T}} = \sum_{j=1}^{q} b_j^{\mathsf{T}} \otimes Y_j^{\mathsf{T}}.$$

Proof. The argument follows [Mal12, Proposition 5.2], with modifications similar to [Sch05, Theorem 2.1] which produce the extra term R_N in the setting of the GOE.

We represent $X_j = \frac{1}{\sqrt{2}}(Z_j + Z_j^\mathsf{T})$ where $Z_j \in \mathbb{R}^{N\times N}$ has i.i.d. $\mathcal{N}(0,1/N)$ entries. The Gaussian integration-by-parts identity $\mathbb{E}[\xi f(\xi)] = N^{-1}\mathbb{E}[f'(\xi)]$ for $\xi \sim \mathcal{N}(0,1/N)$ gives

$$\mathbb{E}\left[(Z_j)_{sl} h_{S_N+T_N}(\Lambda)\right] = \frac{1}{N\sqrt{2}}\mathbb{E}\left[\frac{d}{dt}\Big|_{t=0}(\Lambda\otimes \mathrm{Id}_N - S_N - T_N - t(a_j\otimes e_{sl}^{(N)} + a_j\otimes e_{ls}^{(N)}))^{-1}\right],$$

where $e_{sl}^{(N)} \in \mathbb{R}^{N\times N}$ is the matrix with the single entry (s,l) equal to 1. Applying

$$\frac{d}{dt}A(t)^{-1} = -A(t)^{-1}(t)A'(t)A^{-1}(t), \tag{3.15}$$

we get

$$\mathbb{E}\left[(Z_j)_{sl} h_{S_N+T_N}(\Lambda)\right] = \frac{1}{N\sqrt{2}}\mathbb{E}\left[h_{S_N+T_N}(\Lambda)(a_j\otimes e_{sl}^{(N)} + a_j\otimes e_{ls}^{(N)})h_{S_N+T_N}(\Lambda)\right].$$

Then writing $Z_j = \sum_{s,l=1}^N (Z_j)_{sl} e_{sl}^{(N)}$,

$$\mathbb{E}\left[\left(\frac{a_j}{\sqrt{2}}\otimes Z_j\right)h_{S_N+T_N}(\Lambda)\right] = \frac{1}{2N}\sum_{s,l=1}^N (a_j\otimes e_{sl}^{(N)})\mathbb{E}\left[h_{S_N+T_N}(\Lambda)(a_j\otimes e_{sl}^{(N)} + a_j\otimes e_{ls}^{(N)})h_{S_N+T_N}(\Lambda)\right]. \tag{3.16}$$

For any $a,b\in M_k$ and any elementary tensor $x\otimes Y \in M_k\otimes M_N$,

$$\sum_{s,l=1}^N (a\otimes e_{sl}^{(N)})(x\otimes Y)(b\otimes e_{ls}^{(N)}) = N\,\mathrm{tr}_N(Y)\cdot axb\otimes \mathrm{Id}_N,$$

and

$$\sum_{s,l=1}^N (\mathrm{Id}_k\otimes e_{sl}^{(N)})(x\otimes Y)(\mathrm{Id}_k\otimes e_{sl}^{(N)}) = x\otimes Y^\mathsf{T} = (x^\mathsf{T}\otimes Y)^\mathsf{T} = \sum_{s,l=1}^k (e_{sl}^{(k)}\otimes \mathrm{Id}_N)(x\otimes Y)^\mathsf{T}(e_{sl}^{(k)}\otimes \mathrm{Id}_N).$$

Then by linearity, for any $M\in M_k\otimes M_N$,

$$\sum_{s,l=1}^N (a\otimes e_{sl}^{(N)})M(b\otimes e_{ls}^{(N)}) = N\cdot\Big(a((\mathrm{Id}_k\otimes \mathrm{tr}_N)M)b\Big)\otimes \mathrm{Id}_N,$$

and

$$\sum_{s,l=1}^N (\mathrm{Id}_k\otimes e_{sl}^{(N)})M(\mathrm{Id}_k\otimes e_{sl}^{(N)}) = \sum_{s,l=1}^k (e_{sl}^{(k)}\otimes \mathrm{Id}_N)M^\mathsf{T}(e_{sl}^{(k)}\otimes \mathrm{Id}_N).$$

So the right side of (3.16) is

$$\frac{1}{2}\mathbb{E}\left[(a_j H_{S_N+T_N}(\Lambda)a_j\otimes \mathrm{Id}_N)h_{S_N+T_N}(\Lambda)\right] + \frac{1}{2N}\sum_{s,l=1}^k \mathbb{E}\left[(a_j e_{sl}^{(k)}\otimes \mathrm{Id}_N)h^\mathsf{T}_{S_N+T_N}(\Lambda)(e_{sl}^{(k)}a_j\otimes \mathrm{Id}_N)h_{S_N+T_N}(\Lambda)\right].$$

Applying the same identity as (3.16) for Z_j^T, summing over j, recalling $S_N = \sum_j a_j\otimes (Z_j+Z_j^\mathsf{T})/\sqrt{2}$, multiplying on the left by $Mg_{T_N}(\Gamma)$, and recalling the definition of $\mathcal{R}_s$ from (3.12) we get

$$\begin{aligned}\mathbb{E}\left[Mg_{T_N}(\Gamma)S_N h_{S_N+T_N}(\Lambda)\right] &= \mathbb{E}\left[Mg_{T_N}(\Gamma)(\mathcal{R}_s(H_{S_N+T_N}(\Lambda))\otimes \mathrm{Id}_N)h_{S_N+T_N}(\Lambda)\right]\\ &\quad + \frac{1}{N}\sum_{j=1}^p\sum_{s,l=1}^k \mathbb{E}\left[Mg_{T_N}(\Gamma)(a_j e_{sl}^{(k)}\otimes \mathrm{Id}_N)h^\mathsf{T}_{S_N+T_N}(\Lambda)(e_{sl}^{(k)}a_j\otimes \mathrm{Id}_N)h_{S_N+T_N}(\Lambda)\right].\end{aligned}$$

Writing $S_N = (\Lambda-\Gamma)\otimes \mathrm{Id}_N + (\Gamma\otimes \mathrm{Id}_N - T_N) - (\Lambda\otimes \mathrm{Id}_N - S_N - T_N)$, rearranging, and applying the partial trace $\mathrm{Id}_k\otimes \mathrm{tr}_N$ to both sides yields the result. □

Remark. Proposition 3.6 shows the difference between GOE and GUE matrices. Applying integration by parts for the N^2 independent Gaussian random variables in the GUE setting, we would obtain N^2 terms on the right of (3.16), see [HT05, eqs. (3.7–3.9)]. However, in the GOE setting, there are $2N^2$ terms in (3.16), and the terms in (3.16) which do not appear in the GUE case lead to the first order remainder R_N.

Proposition 3.7. *For any $\Lambda, \Gamma \in M_k^+$ and $M \in M_k \otimes M_N$,*

$$\|R_N(\Lambda, \Gamma, M)\| \leq \frac{C}{N}\|M\|\|(\operatorname{Im}\Lambda)^{-1}\|^2\|(\operatorname{Im}\Gamma)^{-1}\|. \tag{3.17}$$

Proof. This follows from the definition of R_N, and the bounds $\|g_{T_N}(\Gamma)\| \leq \|(\operatorname{Im}\Gamma)^{-1}\|$ and $\|h_{S_N+T_N}(\Lambda)\| \leq \|(\operatorname{Im}\Lambda)^{-1}\|$. □

Proposition 3.8. *For any $\Lambda \in M_k^+$, $M \in M_k \otimes M_N$, and for $\Gamma = \Lambda - \mathcal{R}_s(G_{S_N+T_N}(\Lambda))$,*

$$\|\Theta_N(\Lambda, \Gamma, M)\| \leq \frac{C}{N^2}\|M\|\|(\operatorname{Im}\Lambda)^{-1}\|^5.$$

Proof. The proof is similar to that of [Mal12, Proposition 5.3], and we will omit some details. Introduce

$$K_{S_N+T_N}(\Lambda) = H_{S_N+T_N}(\Lambda) - G_{S_N+T_N}(\Lambda) = H_{S_N+T_N}(\Lambda) - \mathbb{E}[H_{S_N+T_N}(\Lambda)].$$

Then, as $\mathcal{R}_s$ is a linear map, for the given value of Γ

$$\Theta_N(\Lambda, \Gamma, M) = \mathbb{E}\Big[\operatorname{Id}_k \otimes \operatorname{tr}_N\Big[Mg_{T_N}(\Gamma)(\mathcal{R}_s(K_{S_N+T_N}(\Lambda)) \otimes \operatorname{Id}_N)h_{S_N+T_N}(\Lambda)\Big]\Big].$$

Further introduce

$$k_{S_N+T_N}(\Lambda) = h_{S_N+T_N}(\Lambda) - g_{S_N+T_N}(\Lambda) = h_{S_N+T_N}(\Lambda) - \mathbb{E}[h_{S_N+T_N}(\Lambda)].$$

Then, applying $\mathbb{E}[K_{S_N+T_N}(\Lambda)] = 0$, the above implies

$$\Theta_N(\Lambda, \Gamma, M) = \mathbb{E}\Big[\operatorname{Id}_k \otimes \operatorname{tr}_N\Big[Mg_{T_N}(\Gamma)(\mathcal{R}_s(K_{S_N+T_N}(\Lambda)) \otimes \operatorname{Id}_N)k_{S_N+T_N}(\Lambda)\Big]\Big].$$

Denote

$$\|M\|_\infty = \max_{i,j}|M_{ij}|, \quad \|M\|_{\mathrm{HS}}^2 = \sum_{i,j}|M_{ij}|^2.$$

For $X \in M_k \otimes M_N$ and $e_s^{(N)}$ the s^{th} standard basis vector in $\mathbb{C}^N$, define

$$(\operatorname{Id}_k \otimes e_s^{(N)})^{\mathsf{T}} X (\operatorname{Id}_k \otimes e_l^{(N)}) = X^{(s,l)} \in M_k$$

and

$$(e_s^{(k)} \otimes \operatorname{Id}_N)^{\mathsf{T}} X (e_l^{(k)} \otimes \operatorname{Id}_N) = X_{(s,l)} \in M_N.$$

Note in particular that

$$X = \sum_{s,l=1}^N X^{(s,l)} \otimes e_{sl}^{(N)}.$$

Applying this decomposition to $Mg_{T_N}(\Gamma)$ and to $h_{S_N+T_N}(\Lambda)$, we bound

$$\|\Theta(\Lambda, \Gamma, M)\| \leq \sqrt{k}\|\Theta(\Lambda, \Gamma, M)\|_\infty = \frac{\sqrt{k}}{N}\left\|\sum_{s,l=1}^N \mathbb{E}\Big[(Mg_{T_N}(\Gamma))^{(s,l)}\mathcal{R}_s(K_{S_N+T_N}(\Lambda))(k_{S_N+T_N}(\Lambda))^{(l,s)}\Big]\right\|_\infty.$$

Then applying $|\sum_{i=1}^m y_i| \leq m\max_i|y_i|$, we obtain

$$\begin{aligned}
&\|\Theta(\Lambda, \Gamma, M)\| \\
&\leq \frac{k^{5/2}}{N}\max_{u,v,u',v' \in \{1,\ldots,k\}}\left|\mathbb{E}\left[\mathcal{R}_s(K_{S_N+T_N}(\Lambda))_{u',v'}\sum_{s,l=1}^N (Mg_{T_N}(\Gamma))_{u,u'}^{(s,l)} \cdot (k_{S_N+T_N}(\Lambda))_{v',v}^{(l,s)}\right]\right| \\
&\leq k^{5/2}\max_{u,v,u',v' \in \{1,\ldots,k\}}\mathbb{E}\left[|\mathcal{R}_s(K_{S_N+T_N}(\Lambda))_{u',v'}| \cdot \left|\operatorname{tr}_N\left(Mg_{T_N}(\Gamma)_{(u,u')}k_{S_N+T_N}(\Lambda)_{(v',v)}\right)\right|\right] \\
&\leq k^{5/2}\max_{u,v,u',v' \in \{1,\ldots,k\}}\operatorname{Var}[\mathcal{R}_s(H_{S_N+T_N}(\Lambda))_{u',v'}]^{1/2} \cdot \operatorname{Var}\left[\operatorname{tr}_N\left(Mg_{T_N}(\Gamma)_{(u,u')}h_{S_N+T_N}(\Lambda)_{(v',v)}\right)\right]^{1/2},
\end{aligned}$$

where the last line applies Cauchy-Schwarz and Var denotes the complex variance.

Fix any u, v, u', v', and define the scalar-valued functions

$$\begin{aligned}
F_1(S_N) &= \mathcal{R}_s(H_{S_N+T_N}(\Lambda))_{u',v'}, \\
F_2(S_N) &= \operatorname{tr}_N\left(Mg_{T_N}(\Gamma)_{(u,u')}h_{S_N+T_N}(\Lambda)_{(v',v)}\right).
\end{aligned}$$

Following the same arguments as in [Mal12, Proposition 5.3], and applying $\|(\operatorname{Im}\Gamma)^{-1}\| \le \|(\operatorname{Im}\Lambda)^{-1}\|$ because $\Lambda \in M_k^+$ and $\operatorname{Im}\mathcal{R}_s(G_{S_N+T_N}(\Lambda)) \preceq 0$, we may verify that

$$\|\nabla F_1(S_N)\|^2 \le \frac{C}{N}\|(\operatorname{Im}\Lambda)^{-1}\|^4, \qquad \|\nabla F_2(S_N)\|^2 \le \frac{C}{N}\|M\|^2\|(\operatorname{Im}\Lambda)^{-1}\|^6.$$

Then, as the entries of S_N are $C/\sqrt{N}$-Lipschitz in the independent standard Gaussian variables which define $X_1, \dots, X_p$, the Gaussian Poincaré inequality yields

$$\operatorname{Var}[F_1(S_N)] \le \frac{C}{N^2}\|(\operatorname{Im}\Lambda)^{-1}\|^4, \quad \operatorname{Var}[F_2(S_N)] \le \frac{C}{N^2}\|M\|^2\|(\operatorname{Im}\Lambda)^{-1}\|^6.$$

Substituting above concludes the proof. □

Combining Propositions 3.6, 3.7 and 3.8, and specializing to $M = \mathrm{Id}_k \otimes \mathrm{Id}_N$, we obtain Lemma 3.5. The following is then a consequence of the stability property for the subordination equation (3.11), established in [Mal12]: For a parameter $\eta > 0$, define the simply connected open set

$$\Omega_\eta^{(N)} = \left\{\Lambda \in M_k^+ : \|(\operatorname{Im}\Lambda)^{-1}\| < N^\eta\right\}, \tag{3.18}$$

Lemma 3.9 (First-order Cauchy transform approximation)**.** *Let $\eta < 1/3$. Then there exists $N_0 > 0$ such that for all $N \ge N_0$ and $\Lambda \in \Omega_\eta^{(N)}$,*

$$\|G_{s+T_N}(\Lambda) - G_{S_N+T_N}(\Lambda)\| \le \frac{C}{N}\|(\operatorname{Im}\Lambda)^{-1}\|^3\left(1 + \|(\operatorname{Im}\Lambda)^{-1}\|^2\right).$$

Proof. For $\eta < 1/3$, $N \ge N_0$, and $\Lambda \in \Omega_\eta^{(N)}$, Lemma 3.5 implies

$$\|G_{S_N+T_N}(\Lambda) - G_{T_N}(\Lambda - \mathcal{R}_s(G_{S_N+T_N}(\Lambda)))\| \le \frac{C}{N}\|(\operatorname{Im}\Lambda)^{-3}\| \le 1/2.$$

The result then follows from [Mal12, Proposition 4.3]. □

3.3. **Second-order approximation.** For $\Lambda \in M_k^+$, denote the first-order remainder in Lemma 3.5 as

$$R_N(\Lambda) = R_N(\Lambda, \Gamma_N, \mathrm{Id}_k \otimes \mathrm{Id}_N), \qquad \Gamma_N \equiv \Gamma_N(\Lambda) = \Lambda - \mathcal{R}_s(G_{S_N+T_N}(\Lambda)).$$

Define the approximation to Γ_N, which appears in (3.11), by

$$\Gamma_{\mathcal{A}} \equiv \Gamma_{\mathcal{A}}(\Lambda) = \Lambda - \mathcal{R}_s(G_{s+T_N}(\Lambda))).$$

Note that if $\Lambda \in M_k^+$, then $\Gamma_N, \Gamma_{\mathcal{A}} \in M_k^+$ also. Then define an approximation to $R_N(\Lambda)$ by

$$R_{\mathcal{A}}(\Lambda) = -\frac{1}{N}\sum_{j=1}^{p}\sum_{m,l=1}^{k}(\mathrm{Id}_k \otimes \tau_N)\left(g_{T_N}(\Gamma_{\mathcal{A}})(a_j e_{ml}^{(k)} \otimes \mathrm{Id}_N)g_{s+T_N}^{\mathsf{T}}(\Lambda)(e_{ml}^{(k)}a_j \otimes \mathrm{Id}_N)g_{s+T_N}(\Lambda)\right). \tag{3.19}$$

Here, $g_{s+T_N}^{\mathsf{T}} = (\Lambda^{\mathsf{T}} \otimes \mathrm{Id}_N - s^{\mathsf{T}} - T_N^{\mathsf{T}})^{-1}$ where

$$s^{\mathsf{T}} = \sum_{j=1}^{p} a_j^{\mathsf{T}} \otimes x_j$$

and T_N^{T} is as before. In this section, we extend Lemma 3.9 to the following second-order approximation.

Lemma 3.10 (Second-order Cauchy-transform approximation)**.** *For $\eta < 1/3$, a constant $N_0 > 0$, all $N \ge N_0$, and $\Lambda \in \Omega_\eta^{(N)}$,*

$$\|G_{s+T_N}(\Lambda) - G_{S_N+T_N}(\Lambda) + \mathcal{L}_\Lambda(R_{\mathcal{A}}(\Lambda))\| \le \frac{C}{N^2}\|(\operatorname{Im}\Lambda)^{-1}\|^5(1 + \|(\operatorname{Im}\Lambda)^{-1}\|^{10})$$

where $\mathcal{L}_\Lambda : M_k \to M_k$ is the linear map

$$\mathcal{L}_\Lambda(x) = x - G'_{s+T_N}(\Lambda)[\mathcal{R}_s(x)],$$

and $G'_{s+T_N}(\Lambda)$ is the derivative of G_{s+T_N}.

The map $\mathcal{L}_\Lambda$ above appeared also in the analysis of [BC17, Theorem 5.7]. The proof will reveal that $\|\mathcal{L}_\Lambda(R_{\mathcal{A}}(\Lambda))\|$ is of size $O(1/N)$.

We first show that the above result holds with R_N in place of $R_{\mathcal{A}}$.

Proposition 3.11. *For any fixed constant $\eta < 1/3$, there exists $N_0 > 0$ such that for all $N \geq N_0$ and $\Lambda \in \Omega_\eta^{(N)}$,*

$$\|G_{s+T_N}(\Lambda) - G_{S_N+T_N}(\Lambda) + \mathcal{L}_\Lambda(R_N(\Lambda))\| \leq \frac{C}{N^2}\|(\operatorname{Im}\Lambda)^{-1}\|^5(1 + \|(\operatorname{Im}\Lambda)^{-1}\|^{10}).$$

Furthermore, defining the operator norm $\|\mathcal{L}_\Lambda\| = \sup_{x\in M_k:\|x\|=1}\|\mathcal{L}_\Lambda(x)\|$,

$$\|\mathcal{L}_\Lambda\| \leq C(1 + \|(\operatorname{Im}\Lambda)^{-1}\|).$$

Proof. Let us write

$$\Delta_N(\Lambda) = G_{s+T_N}(\Lambda) - G_{S_N+T_N}(\Lambda).$$

Subtracting (3.13) from (3.11), we get

$$\|\Delta_N(\Lambda) - G_{T_N}(\Gamma_\mathcal{A}) + G_{T_N}(\Gamma_N) + R_N(\Lambda)\| \leq \frac{C}{N^2}\|(\operatorname{Im}\Lambda)^{-1}\|^5. \tag{3.20}$$

Lemma 3.9 provides a bound for $\|\Delta_N(\Lambda)\|$, from which we obtain also

$$\|\Gamma_N - \Gamma_\mathcal{A}\| = \|\mathcal{R}_s(\Delta_N(\Lambda))\| \leq \frac{C}{N}\|(\operatorname{Im}\Lambda)^{-1}\|^3\left(1 + \|(\operatorname{Im}\Lambda)^{-1}\|^2\right). \tag{3.21}$$

We apply a Taylor expansion to approximate $G_{T_N}(\Gamma_N) - G_{T_N}(\Gamma_\mathcal{A})$: Fix $v, w \in \mathbb{C}^k$ with $\|v\| = \|w\| = 1$ and define

$$\Gamma_t = (1-t)\Gamma_\mathcal{A} + t\Gamma_N, \qquad f(t) = v^* G_{T_N}(\Gamma_t)w.$$

Then

$$f'(t) = v^*\Big[(\operatorname{Id}\otimes\operatorname{tr}_N)\Big(g_{T_N}(\Gamma_t)((\Gamma_\mathcal{A}-\Gamma_N)\otimes\operatorname{Id}_N)g_{T_N}(\Gamma_t)\Big)\Big]w,$$
$$f''(t) = 2v^*\Big[(\operatorname{Id}\otimes\operatorname{tr}_N)\Big(g_{T_N}(\Gamma_t)((\Gamma_\mathcal{A}-\Gamma_N)\otimes\operatorname{Id}_N)g_{T_N}(\Gamma_t)((\Gamma_\mathcal{A}-\Gamma_N)\otimes\operatorname{Id}_N)g_{T_N}(\Gamma_t)\Big)\Big]w.$$

In particular, for all $t \in [0,1]$, by Lemma 3.9 and the bounds $\|g_{T_N}(\Gamma_t)\| \leq C\|(\operatorname{Im}\Lambda)^{-1}\|$, we find

$$|f''(t)| \leq C\|(\operatorname{Im}\Lambda)^{-1}\|^3\|\Gamma_\mathcal{A} - \Gamma_N\|^2 \leq \frac{C}{N^2}\|(\operatorname{Im}\Lambda)^{-1}\|^9(1 + \|(\operatorname{Im}\Lambda)^{-1}\|^4).$$

So

$$\begin{aligned}\Big|v^*\Big(G_{T_N}(\Gamma_N) - G_{T_N}(\Gamma_\mathcal{A}) - G'_{T_N}(\Gamma_\mathcal{A})[\Gamma_N - \Gamma_\mathcal{A}]\Big)w\Big| &= |f(1) - f(0) - f'(0)| \\ &\leq \frac{C}{N^2}\|(\operatorname{Im}\Lambda)^{-1}\|^9(1 + \|(\operatorname{Im}\Lambda)^{-1}\|^4).\end{aligned}$$

Applying this and $\Gamma_N - \Gamma_\mathcal{A} = \mathcal{R}_s(\Delta_N(\Lambda))$ to (3.20), we obtain

$$\|\Delta_N(\Lambda) + G'_{T_N}(\Gamma_\mathcal{A})[\mathcal{R}_s(\Delta_N(\Lambda))] + R_N(\Lambda)\| \leq \frac{C}{N^2}\|(\operatorname{Im}\Lambda)^{-1}\|^5(1 + \|(\operatorname{Im}\Lambda)^{-1}\|^8). \tag{3.22}$$

We now claim that the linear map

$$F_\Lambda(x) = x + G'_{T_N}(\Gamma_\mathcal{A}(\Lambda))[\mathcal{R}_s(x)]$$

is invertible, with inverse given by $\mathcal{L}_\Lambda$. Indeed, differentiating the subordination identity (3.11) in Λ, for any $\Lambda \in M_k^+$ and $x \in M_k$,

$$G'_{s+T_N}(\Lambda)[x] = G'_{T_N}(\Gamma_\mathcal{A}(\Lambda))\Big[x - \mathcal{R}_s(G'_{s+T_N}(\Lambda)[x])\Big].$$

Then for any $z \in M_k$, setting $x = \mathcal{R}_s(z)$ and $y = z - G'_{s+T_N}(\Lambda)[x] = \mathcal{L}_\Lambda(z)$, we obtain

$$z - y = G'_{T_N}(\Gamma_\mathcal{A})[\mathcal{R}_s(y)].$$

Hence $z = F_\Lambda(y)$, so F_Λ is onto and invertible, with inverse $\mathcal{L}_\Lambda$. Then noting that,

$$F_\Lambda(\Delta_N(\Lambda)) = \Delta_N(\Lambda) + G'_{T_N}(\Gamma_\mathcal{A}(\Lambda))[\mathcal{R}_s(\Delta_N(\Lambda))],$$

we have by (3.22) that

$$\|\Delta_N(\Lambda) + \mathcal{L}_\Lambda(R_N(\Lambda))\| \leq \|\mathcal{L}_\Lambda\| \cdot \|F_\Lambda(\Delta_N(\Lambda)) + R_N(\Lambda)\| \leq \|\mathcal{L}_\Lambda\| \cdot \frac{C}{N^2}\|(\operatorname{Im}\Lambda)^{-1}\|^5(1 + \|(\operatorname{Im}\Lambda)^{-1}\|^8).$$

Finally, writing

$$\begin{aligned}\mathcal{L}_\Lambda(z) &= z - (\mathrm{Id}\otimes\tau_N)\Big[\frac{d}{dt}\Big|_{t=0} g_{s+T_N}(\Lambda + t\mathcal{R}_s(z))\Big] \\ &= z + (\mathrm{Id}\otimes\tau_N)\Big[g_{s+T_N}(\Lambda)(\mathcal{R}_s(z)\otimes \mathrm{Id}_N)g_{s+T_N}(\Lambda)\Big],\end{aligned}$$

we verify $\|\mathcal{L}_\Lambda\| \le C(1+\|(\operatorname{Im}\Lambda)^{-1}\|^2)$, and hence also the desired bound. □

To complete the proof of Lemma 3.10, we will show that

$$\|R_{\mathcal{A}}(\Lambda) - R_N(\Lambda)\| \le \frac{C}{N^2}\|(\operatorname{Im}\Lambda)^{-1}\|^5(1+\|(\operatorname{Im}\Lambda)^{-1}\|^6). \tag{3.23}$$

Let us write

$$R_N(\Lambda) - R_{\mathcal{A}}(\Lambda) = \frac{1}{N}\sum_{j=1}^{p}\sum_{m,l=1}^{k}\left(A_1(j,m,l) + A_2(j,m,l)\right),$$

where

$$A_1(j,m,l) = \mathbb{E}\left[\mathrm{Id}_k\otimes \mathrm{tr}_N\left((g_{T_N}(\Gamma_{\mathcal{A}}) - g_{T_N}(\Gamma_N))(a_j e_{ml}^{(k)}\otimes \mathrm{Id}_N)h^{\mathsf{T}}_{S_N+T_N}(\Lambda)(e_{ml}^{(k)}a_j\otimes \mathrm{Id}_N)h_{S_N+T_N}(\Lambda)\right)\right]$$

and

$$\begin{aligned}A_2(j,m,l) = \mathrm{Id}_k\otimes\tau_N&\left(g_{T_N}(\Gamma_{\mathcal{A}})(a_j e_{ml}^{(k)}\otimes \mathrm{Id}_N)g^{\mathsf{T}}_{s+T_N}(\Lambda)(e_{ml}^{(k)}a_j\otimes \mathrm{Id}_N)g_{s+T_N}(\Lambda)\right) \\ &- \mathbb{E}\left[(\mathrm{Id}_k\otimes \mathrm{tr}_N)\left(g_{T_N}(\Gamma_{\mathcal{A}})(a_j e_{ml}^{(k)}\otimes \mathrm{Id}_N)h^{\mathsf{T}}_{S_N+T_N}(\Lambda)(e_{ml}^{(k)}a_j\otimes \mathrm{Id}_N)h_{S_N+T_N}(\Lambda)\right)\right].\end{aligned} \tag{3.24}$$

We bound separately A_1 and A_2.

Proposition 3.12. *Let $\eta < 1/3$. Then for a constant $N_0 > 0$, all $N \ge N_0$, all $\Lambda \in \Omega_\eta^{(N)}$, and all $j \in \{1,\dots,p\}$ and $m,l \in \{1,\dots,k\}$,*

$$\|A_1(j,m,l)\| \le \frac{C}{N}\|(\operatorname{Im}\Lambda)^{-1}\|^7(1+\|(\operatorname{Im}\Lambda)\|^2).$$

Proof. This follows from (3.21), the bounds $\|h_{S_N+T_N}(\Lambda)\| \le \|(\operatorname{Im}\Lambda)^{-1}\|$ and $g_{T_N}(\Gamma_*) \le \|(\operatorname{Im}\Lambda)^{-1}\|$ for $\Gamma_* \in \{\Gamma_{\mathcal{A}}, \Gamma_N\}$, and the resolvent identity

$$g_{T_N}(\Gamma_{\mathcal{A}}) - g_{T_N}(\Gamma_N) = g_{T_N}(\Gamma_{\mathcal{A}})(\Gamma_N - \Gamma_{\mathcal{A}})g_{T_N}(\Gamma_N).$$ □

To bound $A_2(j,m,l)$, denote by

$$\mathcal{Y}_N = \langle Y_1,\dots,Y_q\rangle$$

the von Neumann subalgebra generated by $Y_1,\dots,Y_q$, both as a subalgebra of M_N and of $\mathcal{A}_N$. For $M \in M_k\otimes\mathcal{Y}_N$, denote

$$G_{M,s+T_N}(\Lambda) = (\mathrm{Id}_k\otimes\tau_N)\Big(Mg_{s+T_N}(\Lambda)\Big), \qquad G_{M,S_N+T_N}(\Lambda) = (\mathrm{Id}_k\otimes \mathrm{tr}_N)\mathbb{E}\Big[Mh_{S_N+T_N}(\Lambda)\Big].$$

Note that these are "left" M_k-valued Cauchy transforms in the sense of Lemma 3.4. We combine the left subordination identity of that lemma with Proposition 3.6, now applied with a general matrix $M \in M_k\otimes\mathcal{Y}_N$ to obtain the following.

Proposition 3.13. *Let $\eta < 1/3$. Then there exists $N_0 > 0$ such that for all $N \ge N_0$, $\Lambda \in \Omega_\eta^{(N)}$, and $M \in M_k\otimes\mathcal{Y}_N$,*

$$\|G_{M,s+T_N}(\Lambda) - G_{M,S_N+T_N}(\Lambda)\| \le \frac{C}{N}\|M\|\|(\operatorname{Im}\Lambda)^{-1}\|^3(1+\|(\operatorname{Im}\Lambda)^{-1}\|^6).$$

Furthermore, let G' be the derivative in Λ and $\|G'(\Lambda)\| = \sup_{x\in M_k:\|x\|=1}\|G'(\Lambda)[x]\|$. Then

$$\|G'_{M,s+T_N}(\Lambda) - G'_{M,S_N+T_N}(\Lambda)\| \le \frac{C}{N}\|M\|\|(\operatorname{Im}\Lambda)^{-1}\|^4(1+\|(\operatorname{Im}\Lambda)^{-1}\|^6).$$

Proof. Applying Lemma 3.4 with $\mathcal{B} = M_k$, $\tau^{\mathcal{B}} = \mathrm{Id}_k \otimes \tau_N$, $m = M$, $t = T_N$, and $b = \Lambda \otimes \mathrm{Id}_N$, we get

$$G_{M,s+T_N}(\Lambda) = \mathrm{Id}_k \otimes \mathrm{tr}_N \left(M g_{T_N}(\Gamma_{\mathcal{A}}) \right)$$

for $\Gamma_{\mathcal{A}} \equiv \Gamma_{\mathcal{A}}(\Lambda) = \Lambda - \mathcal{R}_s(G_{s+T_N}(\Lambda))$ and $\|\Lambda^{-1}\|$ sufficiently small. Since both sides are analytic functions of $\Lambda \in M_k^+$, this must then hold for all $\Lambda \in M_k^+$.

Then applying Proposition 3.6 with this matrix M,

$$\|G_{M,s+T_N}(\Lambda) - G_{M,S_N+T_N}(\Lambda)\| \leq \|\Theta_N(\Lambda, \Gamma_{\mathcal{A}}, M)\| + \|R_N(\Lambda, \Gamma_{\mathcal{A}}, M)\|.$$

By Proposition 3.7, for the first term we have $\|R_N(\Lambda, \Gamma_{\mathcal{A}}, M)\| \leq CN^{-1}\|M\|\|(\mathrm{Im}\,\Lambda)^{-1}\|^3$. By Proposition 3.8, for the second term we have $\|\Theta_N(\Lambda, \Gamma_N, M)\| \leq CN^{-2}\|M\|\|(\mathrm{Im}\,\Lambda)^{-1}\|^5$. Recalling the definition of Θ_N and setting $K_{S_N+T_N}(\Lambda) = H_{S_N+T_N}(\Lambda) - G_{S_N+T_N}(\Lambda)$,

$$\begin{aligned}\|\Theta_N(\Lambda, \Gamma_N, M) - \Theta_N(\Lambda, \Gamma_{\mathcal{A}}, M)\| \leq{}& \mathbb{E}\left[\left\| M(g_{T_N}(\Gamma_N) - g_{T_N}(\Gamma_{\mathcal{A}})) \Big(\mathcal{R}_s(K_{S_N+T_N}(\Lambda)) \otimes \mathrm{Id}_N \Big) h_{S_N+T_N}(\Lambda) \right\| \right] \\ &+ \mathbb{E}\left[\left\| M g_{T_N}(\Gamma_{\mathcal{A}}) \Big((\Gamma_N - \Gamma_{\mathcal{A}}) \otimes \mathrm{Id}_N \Big) h_{S_N+T_N}(\Lambda) \right\| \right].\end{aligned}$$

Applying again (3.21) and the resolvent identity,

$$\|\Theta_N(\Lambda, \Gamma_N, M) - \Theta_N(\Lambda, \Gamma_{\mathcal{A}}, M)\| \leq \frac{C}{N}\|M\|\|(\mathrm{Im}\,\Lambda)^{-1}\|^5 \left(1 + \|(\mathrm{Im}\,\Lambda)^{-1}\|^2\right)^2.$$

Combining the above yields the desired bound on $G_{M,s+T_N} - G_{M,S_N+T_N}$.

For the difference of the derivatives, we apply the Cauchy integral formula. Let $x \in M_k$ with $\|x\| = 1$. Fix $\eta' \in (\eta, 1/3)$. For $r = \|(\mathrm{Im}\,\Lambda)^{-1}\|^{-1}/2$ and any $z \in \mathbb{C}$ with $|z| < r$, note that $\Lambda + zx \in \Omega_{\eta'}^{(N)}$ because

$$\mathrm{Im}(\Lambda + zx) \succeq \mathrm{Im}\,\Lambda - |r|\|x\|\,\mathrm{Id}_k \succeq \left(\|(\mathrm{Im}\,\Lambda)^{-1}\|^{-1} - r \right) \mathrm{Id}_k \succ N^{-\eta'}\,\mathrm{Id}_k\,.$$

Define a path γ by $\gamma(t) = re^{it}$. Then by the Cauchy integral formula applied entrywise to the matrix-valued analytic function $z \mapsto G_*(\Lambda + zx)$,

$$\begin{aligned}\|(G'_{M,s+T_N}(\Lambda) - G'_{M,S_N+T_N}(\Lambda))[x]\| &= \left\| \frac{d}{dz}\Big|_{z=0} (M, G_{s+T_N}(\Lambda + zx) - G_{M,S_N+T_N}(\Lambda + zx)) \right\| \\ &\leq \frac{1}{r} \max_{t \in [0,2\pi]} \left\{ \|G_{M,s+T_N}(\Lambda + \gamma(t)x) - G_{M,S_N+T_N}(\Lambda + \gamma(t)x)\| \right\} \\ &\leq \frac{C}{Nr}\|M\|\|(\mathrm{Im}\,\Lambda)^{-1}\|^3(1 + \|(\mathrm{Im}\,\Lambda)^{-1}\|^6),\end{aligned}$$

where the last inequality comes the first part of the proposition applied to $\Omega_{\eta'}^{(N)}$. As $\eta < \eta' < 1/3$ are arbitrary, replacing η' by η and applying $r^{-1} = 2\|(\mathrm{Im}\,\Lambda)^{-1}\|$, the derivative bound follows. □

We now bound $A_2(j, m, l)$ following an argument similar to [Sch05, Lemma 4.1].

Proposition 3.14. *Let $\eta < 1/3$. Then for a constant $N_0 > 0$, all $N \geq N_0$, $\Lambda \in \Omega_\eta^{(N)}$, and $j \in \{1, \ldots, p\}$ and $m, l \in \{1, \ldots, k\}$,*

$$\|A_2(j, m, l)\| \leq \frac{C}{N}\|(\mathrm{Im}\,\Lambda)^{-1}\|^5(1 + \|(\mathrm{Im}\,\Lambda)^{-1}\|^6).$$

Proof. For $x \in M_k$, consider the embeddings into M_{2k} given by

$$x^{11} = \begin{pmatrix} x & 0 \\ 0 & 0 \end{pmatrix}, \quad x^{12} = \begin{pmatrix} 0 & x \\ 0 & 0 \end{pmatrix}, \quad x^{21} = \begin{pmatrix} 0 & 0 \\ x & 0 \end{pmatrix}, \quad x^{22} = \begin{pmatrix} 0 & 0 \\ 0 & x \end{pmatrix}.$$

In the block decomposition with respect to $M_{2k} \otimes M_N = (M_k \otimes M_N) \oplus (M_k \otimes M_N)$, set

$$\tilde{S}_N = \begin{pmatrix} S_N^\mathsf{T} & 0 \\ 0 & S_N \end{pmatrix} = \sum_{j=1}^{p} ((a_j^\mathsf{T})^{11} + a_j^{22}) \otimes X_j,$$

$$\tilde{T}_N = \begin{pmatrix} T_N^\mathsf{T} & 0 \\ 0 & T_N \end{pmatrix} = \sum_{j=1}^{q} (b_j^\mathsf{T})^{11} \otimes Y_j^\mathsf{T} + b_j^{22} \otimes Y_j.$$

Define $g_{\tilde{T}_N}, h_{\tilde{S}_N+\tilde{T}_N} : M_{2k} \to M_{2k} \otimes M_N$ analogously to g_{T_N} and $h_{S_N+T_N}$.

Define also $\tilde{\Lambda} = (\Lambda^{\mathsf{T}})^{11} + \Lambda^{22}$ and $\tilde{\Gamma} = (\Gamma^{\mathsf{T}})^{11} + \Gamma^{22}$. Note that if $\Lambda, \Gamma \in M_k^+$, then $\tilde{\Lambda}, \tilde{\Gamma} \in M_{2k}^+$. Furthermore, if $\|(\operatorname{Im}\Lambda)^{-1}\| < N^{\eta}$, then $\|(\operatorname{Im}\tilde{\Lambda})^{-1}\| < N^{\eta}$ also. For any $x, y \in M_k$ and $\Lambda, \Gamma \in M_k^+$, we have

$$\begin{aligned}\begin{pmatrix} 0 & 0 \\ 0 & g_{T_N}(\Gamma)(y \otimes \mathrm{Id}_N) h^{\mathsf{T}}_{S_N+T_N}(\Lambda)(x \otimes \mathrm{Id}_N) h_{S_N+T_N}(\Lambda) \end{pmatrix} &= g_{\tilde{T}_N}(\tilde{\Gamma})(y^{21} \otimes \mathrm{Id}_N) h_{\tilde{S}_N+\tilde{T}_N}(\tilde{\Lambda})(x^{12} \otimes \mathrm{Id}_N) h_{\tilde{S}_N+\tilde{T}_N}(\tilde{\Lambda}) \\ &= \frac{d}{dt}\Big|_{t=0} g_{\tilde{T}_N}(\tilde{\Gamma})(y^{21} \otimes \mathrm{Id}_N) h_{\tilde{S}_N+\tilde{T}_N}(\tilde{\Lambda} - t x^{12}).\end{aligned}$$

Therefore,

$$\begin{aligned}&(\mathrm{Id}_k \otimes \mathrm{tr}_N)\Big[g_{T_N}(\Gamma)(y \otimes \mathrm{Id}_N) h^{\mathsf{T}}_{S_N+T_N}(\Lambda)(x \otimes \mathrm{Id}_N) h_{S_N+T_N}(\Lambda)\Big] \\ &= (\mathrm{Tr} \otimes \mathrm{Id}_k)\left[\frac{d}{dt}\Big|_{t=0} (\mathrm{Id}_{2k} \otimes \mathrm{tr}_N)\Big(g_{\tilde{T}_N}(\tilde{\Gamma})(y^{21} \otimes \mathrm{Id}_N) h_{\tilde{S}_N+\tilde{T}_N}(\tilde{\Lambda} - t x^{12})\Big)\right].\end{aligned}$$

We specialize this identity to $\Gamma = \Gamma_{\mathcal{A}}$, $y = a_j e_{ml}^{(k)}$, and $x = e_{ml}^{(k)} a_j$. Set $\tilde{M} = g_{\tilde{T}_N}(\tilde{\Gamma}_{\mathcal{A}})((a_j e_{ml}^{(k)})^{21} \otimes \mathrm{Id}_N)$, and define for $w \in M_{2k}^+$ the left Cauchy transform

$$\tilde{G}_{\tilde{M}, \tilde{S}_N+\tilde{T}_N}(w) = (\mathrm{Id}_{2k} \otimes \mathrm{tr}_N)\mathbb{E}[\tilde{M} h_{\tilde{S}_N+\tilde{T}_N}(w)].$$

Then we obtain that the second term defining $A_2(j, m, l)$ in (3.24) is equal to

$$-\operatorname{Tr} \otimes \mathrm{Id}_k \left[\tilde{G}'_{\tilde{M}, \tilde{S}_N+\tilde{T}_N}(\tilde{\Lambda})[(e_{ml}^{(k)} a_j)^{12}]\right].$$

Similar arguments in the space $M_{2k} \otimes \mathcal{A}_N$ yield that the first term defining $A_2(j, m, l)$ is equal to

$$-\operatorname{Tr} \otimes \mathrm{Id}_k \left[\tilde{G}'_{\tilde{M}, \tilde{s}+\tilde{T}_N}(\tilde{\Lambda})[(e_{ml}^{(k)} a_j)^{12}]\right],$$

where

$$\tilde{s} = \sum_{j=1}^{p} ((a_j^{\mathsf{T}})^{11} + a_j^{22}) \otimes x_j.$$

Taking the difference, we apply Proposition 3.13 with $2k$, $2q$, and $Y_1^{\mathsf{T}}, \ldots, Y_q^{\mathsf{T}}, Y_1, \ldots, Y_q$ in place of k, q, and $Y_1, \ldots, Y_q$. Finally, using the bound $\|\tilde{M}\| \le C\|g_{T_N}(\Gamma_{\mathcal{A}})\| \le \|(\operatorname{Im}\Lambda)^{-1}\|$, we get the desired bound for $A_2(j, m, l)$. □

Combining Propositions 3.12 and 3.14 for A_1 and A_2, we get (3.23). Lemma 3.10 then follows from this and Proposition 3.11.

3.4. **The spectrum of L_N.** Recall the linear polynomials L_N and $L_{\mathcal{A}}$ from (3.9) and (3.10). We now apply Lemma 3.10 to obtain the following spectral inclusion.

Lemma 3.15. *In the setting of Theorem 3.1, for any $k \ge 1$, self-adjoint linear $*$-polynomial L with coefficients in $M_k(\mathbb{C})$, and $\delta > 0$, almost surely for all large N*

$$\operatorname{spec}(L(\mathbf{X}_N, \mathbf{Y}_N)) \subseteq \operatorname{spec}(L(\mathbf{x}, \mathbf{Y}_N))_{\delta}. \tag{3.25}$$

For this, we specialize Lemma 3.10 to the scalar-valued Stieltjes transforms of L_N and $L_{\mathcal{A}}$. For $\lambda \in \mathbb{C}^+$, define

$$\begin{aligned}g_N(\lambda) &= \mathbb{E}[(\mathrm{tr}_k \otimes \mathrm{tr}_N)(\lambda \mathrm{Id}_k \otimes \mathrm{Id}_N - L_N)^{-1}] = \mathrm{tr}_k(G_{S_N+T_N}(\lambda \mathrm{Id}_k - a_0)), \\ g_{\mathcal{A}}(\lambda) &= (\mathrm{tr}_k \otimes \tau_N)(\lambda \mathrm{Id}_k \otimes \mathrm{Id}_N - L_{\mathcal{A}})^{-1} = \mathrm{tr}_k(G_{s+T_N}(\lambda \mathrm{Id}_k - a_0)), \\ r_{\mathcal{A}}(\lambda) &= \mathrm{tr}_k[\mathcal{L}_{(\lambda \mathrm{Id}_k - a_0)}(R_{\mathcal{A}}(\lambda \mathrm{Id}_k - a_0))]\end{aligned}$$

Then Lemma 3.10 applied with $\Lambda = \lambda \mathrm{Id}_k - a_0$ yields

$$\Big|g_{\mathcal{A}}(\lambda) - g_N(\lambda) + r_{\mathcal{A}}(\lambda)\Big| \le \frac{C}{N^2}(\operatorname{Im}\lambda)^{-5}(1 + (\operatorname{Im}\lambda)^{-10}) \tag{3.26}$$

for any $\eta \in (0, 1/3)$, a constant $C \equiv C(\eta) > 0$, and all $\lambda \in \mathbb{C}^+$ such that $\operatorname{Im}\lambda \ge N^{-\eta}$.

As in [Sch05], we first show the following.

Proposition 3.16. *The function $r_{\mathcal{A}}(\lambda)$ is the Stieltjes transform of a distribution on $\mathbb{R}$ with support contained in* $\operatorname{spec}(L_{\mathcal{A}})$.

Proof. By [Sch05, Theorem 5.4], it suffices to check that

- $r_{\mathcal{A}}(\lambda)$ is analytic on $\mathbb{C} \setminus \operatorname{spec}(L_{\mathcal{A}})$,
- $r_{\mathcal{A}}(\lambda) \to 0$ as $|\lambda| \to \infty$, and
- There exists a constant $C > 0$ and a compact set $K \subset \mathbb{R}$ containing $\operatorname{spec}(L_{\mathcal{A}})$ such that $|r_{\mathcal{A}}(\lambda)| \leq C \cdot \max\{\operatorname{dist}(\lambda, K)^{-3}, 1\}$ for all $\lambda \in \mathbb{C} \setminus \mathbb{R}$.

The matrix $\Gamma_{\mathcal{A}}$ in (3.19) is given by $\Gamma_{\mathcal{A}}(\lambda) = \lambda \operatorname{Id}_k - a_0 - \mathcal{R}_s(G_{s+T_N}(\lambda \operatorname{Id}_k - a_0))$. For the first claim, if $\lambda \notin \operatorname{spec}(L_{\mathcal{A}})$, then $G_{s+T_N}(\lambda \operatorname{Id}_k - a_0)$ exists and is analytic at λ. The subordination identity (3.11) implies $G_{s+T_N}(\lambda \operatorname{Id}_k - a_0) = G_{T_N}(\Gamma_{\mathcal{A}}(\lambda))$ for all $\lambda \in \mathbb{C}^+$, and hence also for all $\lambda \notin \operatorname{spec}(L_{\mathcal{A}})$ by analytic continuation. Then $g_{T_N}(\Gamma_{\mathcal{A}}(\lambda))$ also exists and is analytic at λ. Recalling the definition of $r_{\mathcal{A}}$ above and of $R_{\mathcal{A}}$ from (3.19), we see that $r_{\mathcal{A}}(\lambda)$ is analytic on $\mathbb{C} \setminus \operatorname{spec}(L_{\mathcal{A}})$.

For the second claim, note that for some constant $M > 0$, uniformly over $\lambda \in \mathbb{C}$ where $|\lambda| \geq M$, we have

$$\|G_{s+T_N}(\lambda \operatorname{Id}_k - a_0)\| \leq \|(\lambda \operatorname{Id}_k \otimes \operatorname{Id}_N - L_{\mathcal{A}})^{-1}\| \leq C/|\lambda|$$

and similarly $\|G'_{s+T_N}(\lambda \operatorname{Id}_k - a_0)\| \leq C/|\lambda|^2$. Then also

$$\|G_{T_N}(\Gamma_A(\lambda))\| \leq \frac{1}{|\lambda| - \|a_0\| - \|\mathcal{R}_s(G_{s+T_N}(\lambda \operatorname{Id}_k - a_0))\| - \|T_N\|} \leq C/|\lambda|.$$

Thus $\|R_{\mathcal{A}}(\lambda \operatorname{Id}_k - a_0)\| \leq C|\lambda|^{-3}$, and $|r_{\mathcal{A}}(\lambda)| \leq C|\lambda|^{-3}(1 + |\lambda|^{-2})$. In particular, $r_{\mathcal{A}}(\lambda) \to 0$ as $|\lambda| \to \infty$

For the third claim, let $K = [-M, M]$. Over the region $\operatorname{Re}\lambda \in K$ and $\operatorname{Im}\lambda \in [-M, M] \setminus \{0\}$, we apply the similar bound

$$|r_{\mathcal{A}}(\lambda)| \leq C|\operatorname{Im}\lambda|^{-3}(1 + |\operatorname{Im}\lambda|^{-2})$$

to get $|r_{\mathcal{A}}(\lambda)| \leq C \operatorname{dist}(\lambda, K)^{-3}$. For λ outside this region, the preceding argument implies $|r_{\mathcal{A}}(\lambda)|$ is uniformly bounded. The third claim follows. □

Combining this with (3.26), we get the following result.

Lemma 3.17. *Fix any $M, \delta > 0$ such that $\operatorname{spec}(L_{\mathcal{A}})_\delta \subset [-M, M]$ for all large N. Consider any (sequence of) non-negative smooth functions $f_N : \mathbb{R} \to [0, 1]$ such that*

$$f_N(x) = \begin{cases} 0 & x \in \operatorname{spec}(L_{\mathcal{A}})_{\delta/2} \text{ or } x \notin [-M - \delta, M + \delta] \\ 1 & x \in [-M, M] \setminus \operatorname{spec}(L_{\mathcal{A}})_\delta \end{cases}$$

and $|f_N^{(k)}(x)| \leq C_k \delta^{-k}$ for each $k \geq 1$, some constants $C_k > 0$, and all $x \in \mathbb{R}$. Then for any fixed $\kappa \in (0, 1/2)$, almost surely as $N \to \infty$,

$$N^{1+\kappa}(\operatorname{tr}_k \otimes \operatorname{tr}_N)(f_N(L_N)) \to 0.$$

Proof. The argument is similar to [HT05] and [Mal12], and we will omit most of the details. Since $f_N \equiv 0$ on $\operatorname{spec}(L_{\mathcal{A}})$, we have from Proposition 3.16 and the Stieltjes inversion formula that

$$\begin{aligned} \mathbb{E}[(\operatorname{tr}_k \otimes \operatorname{tr}_N) f_N(L_N)] &= \lim_{y \to 0^+} -\frac{1}{\pi} \operatorname{Im}\left[\int_{\mathbb{R}} f_N(x) g_N(x + iy)] dx\right] \\ &= \lim_{y \to 0^+} \frac{1}{\pi} \operatorname{Im}\left[\int_{\mathbb{R}} f_N(x)[g_{\mathcal{A}}(x + iy) + r_{\mathcal{A}}(x + iy) - g_N(x + iy)] dx\right]. \end{aligned}$$

Then applying (3.26) and following the same arguments as [HT05, Theorem 6.2], we get

$$\mathbb{E}[(\operatorname{tr}_k \otimes \operatorname{tr}_N) f_N(L_N)] \leq C/N^2$$

for a constant $C \equiv C(\delta) > 0$.

As in the proof of Proposition 3.6, we write $X_j = \frac{1}{\sqrt{2}}(Z_j + Z_j^{\mathsf{T}})$ where $Z_j \in \mathbb{R}^{N \times N}$ has i.i.d. $\mathcal{N}(0, 1/N)$ entries. Defining

$$F_N(Z_1, ..., Z_p) = f_N\left(a_0 \otimes \operatorname{Id}_N + \frac{1}{\sqrt{2}} \sum_{j=1}^{p} a_j \otimes (Z_j + Z_j^{\mathsf{T}}) + \sum_{j=1}^{q} b_j \otimes Y_j\right),$$

the Gaussian Poincaré inequality yields

$$\operatorname{Var}[(\operatorname{tr}_k \otimes \operatorname{tr}_N) f_N(L_N)] \leq \frac{1}{N} \mathbb{E}\left[\|\nabla F_N(Z_1, ..., Z_p)\|_2^2\right].$$

The same argument as [HT05, Proposition 4.7] yields

$$\|\nabla F_N(Z_1,...,Z_p)\|_2^2 \le \frac{C}{N}(\mathrm{tr}_k \otimes \mathrm{tr}_N)((f_N')^2(L_N)),$$

where $(f_N')^2$ denotes the function $z \mapsto (f_N'(z))^2$. So

$$\mathrm{Var}[(\mathrm{tr}_k \otimes \mathrm{tr}_N) f(L_N)] \le \frac{C}{N^2}\mathbb{E}[(\mathrm{tr}_k \otimes \mathrm{tr}_N)((f_N')^2(L_N))].$$

Applying the same argument as above,

$$\mathbb{E}[(\mathrm{tr}_k \otimes \mathrm{tr}_N)(f_N')^2(L_N)] = \lim_{y\to 0^+}\frac{1}{\pi}\,\mathrm{Im}\left[\int_{\mathbb{R}}(f_N'(x))^2[g_{\mathcal{A}}(x+iy)+r_{\mathcal{A}}(x+iy)-g_N(x+iy)]dx\right] \le C/N^2,$$

so $\mathrm{Var}[(\mathrm{tr}_k \otimes \mathrm{tr}_N) f_N(L_N)] \le C/N^4$. Then by Markov's inequality,

$$\mathbb{P}[(\mathrm{tr}_k \otimes \mathrm{tr}_N) f_N(L_N) \ge N^{-1-\kappa}] \le N^{2+2\kappa}\mathbb{E}[((\mathrm{tr}_k \otimes \mathrm{tr}_N) f_N(L_N))^2] \le CN^{2+2\kappa}\cdot N^{-4}.$$

Taking $0 < \kappa < 1/2$, the result follows from Borel-Cantelli. □

Taking a constant $M > 0$ large enough ensures $\mathrm{spec}(L_N) \subset [-M, M]$ almost surely for all large N. Then defining f_N as in Lemma 3.17, if there exists an eigenvalue of L_N outside $\mathrm{spec}(L_{\mathcal{A}})_\delta$, we must have

$$(\tau_k \otimes \tau_N) f_N(L_N) \ge N^{-1}.$$

Lemma 3.17 guarantees this does not happen, almost surely for all large N. This concludes the proof of Lemma 3.15.

3.5. **Linearization trick and ultraproduct argument.** We conclude the proof of Theorem 3.1 from Lemma 3.15 by applying the linearization trick and ultraproduct argument of [HT05]. As our algebra $\mathcal{A}$ is N-dependent, we apply this argument in a subsequence form.

Let $M_k(\mathbb{Q}+i\mathbb{Q})_{sa}$ be the set of $k \times k$ Hermitian matrices whose entries have rational real and imaginary parts. Define the countable set

$$\mathcal{L} = \bigcup_{k=1}^{\infty}\{\text{all linear } *\text{-polynomials of } p+q \text{ variables with coefficients in } M_k(\mathbb{Q}+i\mathbb{Q})_{sa}\}.$$

Let Ω denote the sample space. Let $\mathbf{Z}_N(\omega) = (\mathbf{X}_N(\omega), \mathbf{Y}_N)$, $\mathbf{z}_N = (\mathbf{x}, \mathbf{Y}_N)$, for all $\omega \in \Omega$.

Proof of Theorem 3.1. Let $\Omega_0 \subset \Omega$ be the event where

$$\sup_{N\ge 1}\max_{i=1}^{p}\left\|X_i^{(N)}(\omega)\right\| < \infty,$$

and also where for each $L \in \mathcal{L}$ and (rational) $\delta > 0$, there exists $N_0(L,\delta,\omega) > 0$ such that

$$\mathrm{spec}(L(\mathbf{Z}_N(\omega))) \subseteq \mathrm{spec}(L(\mathbf{z}_N))_\delta \tag{3.27}$$

for all $N \ge N_0(L,\delta,\omega)$. By Lemma 3.15, Ω_0 has probability 1.

We claim that (3.1) holds on Ω_0. Suppose by contradiction that this is false for some non-commutative $*$-polynomial Q (with coefficients in $\mathbb{C}$), $\delta > 0$, and $\omega \in \Omega_0$. Then at this ω, there is a subsequence $\{N_j\}$ and values $\{\lambda_{N_j}\} \in \mathbb{R}$ such that for all j,

$$\lambda_{N_j} \in \mathrm{spec}(Q(\mathbf{Z}_{N_j}(\omega))) \quad \text{but} \quad \lambda_{N_j} \notin \mathrm{spec}(Q(\mathbf{z}_{N_j}))_\delta. \tag{3.28}$$

Since $\mathrm{spec}(Q(\mathbf{Z}_N(\omega)))$ is uniformly bounded in N, there is a further subsequence $\{N_{j_m}\}$ such that (3.28) still holds and

$$\lambda_{N_{j_m}} \to \lambda_0 \quad \text{as} \quad N_{j_m} \to \infty, \tag{3.29}$$

for some $\lambda_0 \in \mathbb{R}$. To ease notation, let us denote $\{N_{j_m}\}$ in the following argument simply as $\{N\}$.

We introduce the quotient map defined in [HT05, Proposition 7.3]. Define the product and sum of the sequence of algebras $\{\mathcal{A}_N\}_{N=1}^{\infty}$ by

$$\prod_N \mathcal{A}_N = \left\{(a_N)_{N=1}^{\infty} : a_N \in \mathcal{A}_N, \sup_N \|a_N\| < \infty\right\}$$

and

$$\sum_N \mathcal{A}_N = \left\{(a_N)_{N=1}^{\infty} : a_N \in \mathcal{A}_N,\ \lim_{N\to\infty} \|a_N\| = 0\right\}.$$

Then $\prod_N \mathcal{A}_N$ is a C^*-algebra (under coordinate-wise addition and multiplication), and $\sum_N \mathcal{A}_N$ is a two-sided ideal. Thus, we can define a quotient map by

$$\pi_{\mathcal{A}} : \prod_N \mathcal{A}_N \longrightarrow \Big(\prod_N \mathcal{A}_N\Big)\Big/\Big(\sum_N \mathcal{A}_N\Big) \equiv \mathcal{C}_{\mathcal{A}}.$$

We identify $M_k \otimes \mathcal{C}_{\mathcal{A}}$ with

$$\Big(\prod_N M_k \otimes \mathcal{A}_N\Big)\Big/\Big(\sum_N M_k \otimes \mathcal{A}_N\Big).$$

Similarly, define the product and sum of the matrix spaces $\{M_N\}_{N=1}^{\infty}$, and a quotient map

$$\pi : \prod_N M_N \longrightarrow \Big(\prod_N M_N\Big)\Big/\Big(\sum_N M_N\Big) \equiv \mathcal{C}.$$

Denote $\mathbf{Z}_N(\omega) = (\mathbf{X}_N(\omega), \mathbf{Y}_N)$ and $\mathbf{z}_N = (\mathbf{x}, \mathbf{Y}_N)$. Denote their images under the above quotient maps as

$$\mathbf{Z}'(\omega) = (Z_i'(\omega))_{i=1}^{p+q} = \left(\pi\left(\left\{X_1^{(N)}(\omega)\right\}\right), \ldots, \pi\left(\left\{X_p^{(N)}(\omega)\right\}\right), \pi\left(\left\{Y_1^{(N)}\right\}\right), \ldots, \pi\left(\left\{Y_q^{(N)}\right\}\right)\right),$$

$$\mathbf{z}' = (z_i')_{i=1}^{p+q} = \left(\pi_{\mathcal{A}}\left(\{x_1\}\right), \ldots, \pi_{\mathcal{A}}\left(\{x_p\}\right), \pi_{\mathcal{A}}\left(\left\{Y_1^{(N)}\right\}\right), \ldots, \pi_{\mathcal{A}}\left(\left\{Y_q^{(N)}\right\}\right)\right).$$

We first claim that for every $L \in \mathcal{L}$,

$$\operatorname{spec}(L(\mathbf{Z}'(\omega))) \subseteq \operatorname{spec}(L(\mathbf{z}')). \tag{3.30}$$

Indeed, fixing $L \in \mathcal{L}$, for any $\lambda \notin \operatorname{spec}(L(\mathbf{z}'))$ there exists an element $w' \in M_k \otimes \mathcal{C}_{\mathcal{A}}$ such that $w'(\lambda - L(\mathbf{z}')) = 1$. Letting $(w_N)_{N=1}^{\infty} \in \prod_N M_k \otimes \mathcal{A}_N$ be such that $\pi_{\mathcal{A}}(\{w_N\}) = w'$, and noting that $\operatorname{Id}_k \otimes \pi_{\mathcal{A}}(\{L(\mathbf{z}_N)\}) = L(\mathbf{z}')$, there must exist $(v_N)_{N=1}^{\infty} \in \sum_N M_k \otimes \mathcal{A}_N$ such that for every N,

$$w_N\left(\lambda \operatorname{Id}_k \otimes \operatorname{Id}_N - L(\mathbf{z}_N)\right) = \operatorname{Id}_k \otimes \operatorname{Id}_N + v_N.$$

For N large enough such that $\|v_N\| < 1/2$, we get

$$\lambda \notin \operatorname{spec}(L(\mathbf{z}_N)) \quad \text{and} \quad \left\|\left(\lambda \operatorname{Id}_k \otimes \operatorname{Id}_N - L(\mathbf{z}_N)\right)^{-1}\right\| \leq 2 \sup_N \|w_N\|.$$

Then $\operatorname{dist}(\lambda, \operatorname{spec}(L(\mathbf{z}_N))) \geq (2\sup_N \|w_N\|)^{-1}$. Applying (3.27) with $\delta = (4\sup_N \|w_N\|)^{-1}$, we conclude that $\operatorname{dist}(\lambda, \operatorname{spec}(L(\mathbf{Z}_N(\omega)))) \geq (4\sup_N \|w_N\|)^{-1}$, so $\lambda \notin \operatorname{spec}(L(\mathbf{Z}_N(\omega)))$ for all large N. Then defining $\{W_N\}_{N=1}^{\infty}$ by $(\lambda \operatorname{Id}_k \otimes \operatorname{Id}_N - L(\mathbf{Z}_N(\omega)))^{-1}$ for large N, we obtain that $\pi(\{W_N\})$ is the inverse of $\lambda - L(\mathbf{Z}'(\omega))$ in $M_k \otimes \mathcal{C}$. Thus, $\lambda \notin \operatorname{spec}(L(\mathbf{Z}'(\omega)))$, so (3.30) holds.

Then for this fixed ω, [HT05, Theorem 2.4] establishes the existence of a unital $*$-homomorphism

$$\phi : \langle 1, z_1', \ldots, z_{p+q}' \rangle \to \langle 1, Z_1'(\omega), \ldots, Z_{p+q}'(\omega) \rangle$$

such that $\phi(z_i') = Z_i'(\omega)$ for each $i \in \{1, \ldots, p+q\}$. Note that if $x \in \langle 1, z_1', \ldots, z_{p+q}' \rangle$ is invertible with inverse x^{-1}, then $\phi(x)$ is also invertible with inverse $\phi(x^{-1})$. The assumption (3.28) implies that $\operatorname{dist}(\lambda_N, \operatorname{spec}(Q(\mathbf{z}_N))) \geq \delta$ for all N. Then

$$\pi_{\mathcal{A}}\left(\{\lambda_N \operatorname{Id}_N - Q(\mathbf{z}_N)\}_{N=1}^{\infty}\right) = \pi_{\mathcal{A}}(\{\lambda_N \operatorname{Id}_N\}) - Q(\mathbf{z}')$$

is invertible. Applying $\phi(Q(\mathbf{z}')) = Q(\mathbf{Z}'(\omega))$, we get that $\phi(\pi_{\mathcal{A}}(\{\lambda_N \operatorname{Id}_N\})) - Q(\mathbf{Z}'(\omega))$ is also invertible. From (3.29), we obtain

$$\pi(\{\lambda_N \operatorname{Id}_N\}) = \pi(\{\lambda_0 \operatorname{Id}_N\}) = \lambda_0 \mathbf{1}_{\mathcal{C}} = \phi(\lambda_0 \mathbf{1}_{\mathcal{C}_{\mathcal{A}}}) = \phi(\pi_{\mathcal{A}}(\{\lambda_0 \operatorname{Id}_N\})) = \phi(\pi_{\mathcal{A}}(\{\lambda_N \operatorname{Id}_N\})). \tag{3.31}$$

Then $\pi(\{\lambda_N \operatorname{Id}_N - Q(\mathbf{Z}_N(\omega))\}_{N=1}^{\infty})$ is invertible. So $W_N(\lambda_N \operatorname{Id}_N - Q(\mathbf{Z}_N(\omega))) = \operatorname{Id}_N + V_N$ for some matrices W_N, V_N with $\sup_N \|W_N\| < \infty$ and $\|V_N\| \to 0$. For large enough N, this contradicts the first statement of (3.28), that $\lambda_N \in \operatorname{spec}(Q(\mathbf{Z}_N))$, concluding the proof. □

Proof of Theorem 3.2. The convergence in trace in (3.2) is known, see e.g. [AGZ10, Theorem 5.4.5]. To verify the convergence in norm, it is sufficient to show almost surely

$$\liminf_{N\to\infty} \|Q(\mathbf{X}_N, \mathbf{Y}_N)\| \geq \|Q(\mathbf{x}, \mathbf{y})\|,$$
$$\limsup_{N\to\infty} \|Q(\mathbf{X}_N, \mathbf{Y}_N)\| \leq \|Q(\mathbf{x}, \mathbf{y})\|.$$

The first inequality can be verified from the trace convergence and [HT05, Lemma 7.2]. For the second inequality, because of the linearization trick, it suffices to prove that for any linear polynomial L with coefficients in M_k, any $\delta > 0$, and all large N,

$$\operatorname{spec}(L(\mathbf{X}_N, \mathbf{Y}_N)) \subset \operatorname{spec}(L(\mathbf{x}, \mathbf{y}))_\delta.$$

Based on Lemma 3.15, it remains to show

$$\operatorname{spec}(L(\mathbf{x}, \mathbf{Y}_N)) \subset \operatorname{spec}(L(\mathbf{x}, \mathbf{y}))_\delta,$$

which is the main result in [Mal12, Section 7 and Appendix A]. □

4. Anisotropic resolvent law from free deterministic equivalents

We now describe how deterministic resolvent approximations may be derived in the free deterministic equivalent framework of [SV12]. We consider the following setting for rectangular free deterministic equivalents, described in more detail in [SV12] and [FJ16, Section 3]. Note that taking $k = 1$ yields results for the simpler square setting and non-amalgamated freeness.

4.1. Statement of the result.

Let $\mathcal{A}_1 = \mathbb{C}^{N\times N}$ and $\tau_1 = N^{-1}\operatorname{Tr}$. Let $N = N_1 + \ldots + N_k$, and consider the associated $k \times k$ block decomposition of $\mathcal{A}_1$. Define $P_1, \ldots, P_k \in \mathcal{A}_1$ by

$$P_r = \operatorname{diag}(0, \ldots, 0, \operatorname{Id}_{N_r}, 0, \ldots, 0).$$

These are mutually orthogonal projections summing to Id_N. Let $\mathcal{D} \subset \mathcal{A}_1$ be the subalgebra generated by $P_1, \ldots, P_k$, which is explicitly given by

$$\mathcal{D} = \{z_1 P_1 + \ldots + z_k P_k : z_1, \ldots, z_k \in \mathbb{C}\}.$$

Define the space of block-orthogonal matrices

$$\mathcal{O} = \{\operatorname{diag}(O_1, \ldots, O_k) : O_r \in \mathbb{R}^{N_r\times N_r},\ O_r^{\mathsf{T}} O_r = \operatorname{Id} \text{ for each } r\}.$$

Let $H_1, \ldots, H_p, B_1, \ldots, B_q \in \mathcal{A}_1$ be matrices where

- $H_1, \ldots, H_p$ are deterministic.
- $B_1, \ldots, B_q$ are random and invariant under conjugation by $\mathcal{O}$. That is, $(B_1, \ldots, B_q)$ has the same joint law as $(OB_1O^{\mathsf{T}}, \ldots, OB_qO^{\mathsf{T}})$ for each fixed $O \in \mathcal{O}$.

Consider a general self-adjoint $*$-polynomial Q of $p+q$ arguments, with coefficients in $\mathcal{D}$. We will approximate the resolvent of the Hermitian matrix

$$W = Q(H_1, \ldots, H_p, B_1, \ldots, B_q) \in \mathbb{C}^{N\times N}.$$

Let $(\mathcal{A}_2, \tau_2)$ be a (possibly N-dependent) von Neumann probability space, also containing $\mathcal{D}$ as a subalgebra. Suppose $\mathcal{A}_2$ has elements $b_1, \ldots, b_q$ satisfying:

Assumption 4.1. For any fixed $*$-polynomial Q in q arguments, with coefficients in $\mathcal{D}$, almost surely as $N \to \infty$,

$$\frac{1}{N}\operatorname{Tr}\Big(Q(B_1, \ldots, B_q)\Big) - \tau_2\Big(Q(b_1, \ldots, b_q)\Big) \to 0. \tag{4.1}$$

Define the von Neumann amalgamated free product over $\mathcal{D}$,

$$(\mathcal{A}, \tau) = (\mathcal{A}_1, \tau_1) *_{\mathcal{D}} (\mathcal{A}_2, \tau_2).$$

Then $\{H_1, \ldots, H_p\}, \{b_1, \ldots, b_q\} \in \mathcal{A}$ form a free deterministic equivalent for the matrices $\{H_1, \ldots, H_p\}$ and $\{B_1, \ldots, B_q\}$, in the sense of [FJ16, Definition 3.8]. Set

$$w = Q(H_1, \ldots, H_p, b_1, \ldots, b_q).$$

Taking $N \to \infty$ such that $c < N_r/N < C$ for constants $C, c > 0$, [FJ16, Theorem 3.9] shows that the τ-distribution of w asymptotically approximates the spectral distribution of W.

We extend this here to a deterministic approximation $R_0(z) \in \mathbb{C}^{N\times N}$ of the resolvent $(W - z\,\mathrm{Id})^{-1}$. Let

$$\mathcal{H} = \langle H_1, \dots, H_p, \mathcal{D}\rangle$$

be the generated von Neumann subalgebra of $\mathcal{A}$. Let $\tau^{\mathcal{H}} : \mathcal{A} \to \mathcal{H}$ be the (unique) τ-invariant conditional expectation onto $\mathcal{H}$, which satisfies $\tau \circ \tau^{\mathcal{H}} = \tau$. Importantly, note in particular that for any $a \in \mathcal{A}$,

$$\tau^{\mathcal{H}}(a) \in \mathcal{H} \subset \mathcal{A}_1 \equiv \mathbb{C}^{N\times N},$$

so that $\tau^{\mathcal{H}}(a)$ may be interpreted as an $N \times N$ matrix. We define $R_0(z) = \tau^{\mathcal{H}}((w-z)^{-1})$.

Theorem 4.2 (Anisotropic resolvent law)**.** *Fix constants $C, c, \delta > 0$ and a self-adjoint $*$-polynomial Q. Let W, w, and $\mathcal{H} \subset \mathcal{A}$ be as defined above, where $\{B_j\}$ and $\{b_j\}$ satisfy Assumption 4.1. Suppose in addition that $c < N_r/N < C$, $\|H_i\| < C$, and $\|B_j\| < C$ for all r, i, j, almost surely for all large N. Set*

$$R_0(z) = \tau^{\mathcal{H}}((w-z)^{-1}), \qquad \mathbb{D} = \{z \in \mathbb{C} : \mathrm{dist}(z, \mathrm{spec}(w)) \geq \delta \text{ and } \mathrm{dist}(z, \mathrm{spec}(W)) \geq \delta\}. \tag{4.2}$$

Then for any (sequence of) deterministic unit vectors $u, v \in \mathbb{C}^N$, almost surely as $N \to \infty$,

$$\sup_{z\in\mathbb{D}} |u^*(W - z\,\mathrm{Id})^{-1}v - u^*R_0(z)v| \to 0. \tag{4.3}$$

In applications with rectangular matrices, $k \geq 2$, we are typically interested in self-adjoint $*$-polynomials Q which have only the $(1,1)$-block nonzero. That is, W and w satisfy

$$W = P_1 W P_1, \qquad w = P_1 w P_1.$$

In this setting, since $\mathcal{D} \subset \mathcal{H}$, we have $P_r R_0(z) P_s = \tau^{\mathcal{H}}(P_r(w-z)^{-1}P_s)$. Then outside the $(1,1)$-block, the resolvent approximation $R_0(z)$ has the structure

$$P_r R_0(z) P_s = 0 \text{ for all } r \neq s, \qquad P_r R_0(z) P_r = -z^{-1} P_r \text{ for all } r \neq 1.$$

Denote by $W_{11} \in \mathbb{C}^{N_1\times N_1}$ the $(1,1)$-block of W. Analogous to $\mathbb{C}^{N_1\times N_1}$ is a "compressed algebra" $\mathcal{A}^c = \{P_1 a P_1 : a \in \mathcal{A}\}$ with unit P_1 [SV12]. Denote by $w_{11} \in \mathcal{A}^c$ and $\mathrm{spec}(w_{11})$ the element w and its spectrum, viewed as a self-adjoint operator in $\mathcal{A}^c$.

Corollary 4.3. *In the setting of Theorem 4.2, suppose in addition that $W = P_1 W P_1$ and $w = P_1 w P_1$, and let W_{11} and w_{11} be as above. Let $(R_0(z))_{11} \in \mathbb{C}^{N_1\times N_1}$ be the $(1,1)$-block of $R_0(z) = \tau^{\mathcal{H}}((w-z)^{-1})$, and set*

$$\mathbb{D}_1 = \{z \in \mathbb{C} : \mathrm{dist}(z, \mathrm{spec}(w_{11})) \geq \delta \text{ and } \mathrm{dist}(z, \mathrm{spec}(W_{11})) \geq \delta\}.$$

Then for any (sequence of) deterministic unit vectors $u_1, v_1 \in \mathbb{C}^{N_1}$, almost surely as $N \to \infty$,

$$\sup_{z\in\mathbb{D}_1} |u_1^*(W_{11} - z\,\mathrm{Id})^{-1}v_1 - u_1^*(R_0(z))_{11}v_1| \to 0. \tag{4.4}$$

In the remainder of this section, we prove the above results.

4.2. **Convergence for moments.** To prove Theorem 4.2, we require first the following result showing convergence of moments.

Theorem 4.4. *Under the assumptions of Theorem 4.2, let Q be any fixed $*$-polynomial of $p+q$ arguments, with coefficients in $\mathcal{D}$, and $v, w \in \mathbb{C}^N$ any deterministic vectors such that $\|v\|, \|w\| \leq C$. Then almost surely as $N \to \infty$,*

$$v^*Q(H_1, \dots, H_p, B_1, \dots, B_q)w - v^*\tau^{\mathcal{H}}(Q(H_1, \dots, H_p, b_1, \dots, b_q))w \to 0.$$

Call a matrix $A \in \mathbb{C}^{N\times N}$ (or element $a \in \mathcal{A}$) simple if $P_r A P_s = A$ (resp. $P_r a P_s = a$) for some $r, s \in \{1, \dots, k\}$. By linearity, we may reduce Theorem 4.4 to the following setting.

Lemma 4.5. *Fix the constants $C, c > 0$. Suppose, in addition to the assumptions of Theorem 4.4, that each H_i, B_j, and b_j is simple for $i = 1, \dots, p$ and $j = 1, \dots, q$. Then for any $m \geq 0$, any $j_1, \dots, j_m \in \{1, \dots, q\}$ and $\{i_1, \dots, i_{m-1}\} \in \{1, \dots, p\}$, and any deterministic $v, w \in \mathbb{C}^N$ with $\|v\|, \|w\| \leq C$, almost surely as $N \to \infty$,*

$$v^*B_{j_1}H_{i_1}\dots B_{j_{m-1}}H_{i_{m-1}}B_{j_m}w - v^*\tau^{\mathcal{H}}(b_{j_1}H_{i_1}\dots b_{j_{m-1}}H_{i_{m-1}}b_{j_m})w \to 0. \tag{4.5}$$

We first explain why Theorem 4.4 follows, and then prove the lemma by induction on m.

Proof of Theorem 4.4. Any $A \in \mathbb{C}^{N\times N}$ or $a \in \mathcal{A}$ is decomposed into simple elements as

$$A = \sum_{r=1}^{k}\sum_{s=1}^{k} P_r A P_s, \qquad a = \sum_{r=1}^{k}\sum_{s=1}^{k} p_r a p_s.$$

Then by linearity, it suffices to establish Theorem 4.4 for all $*$-monomials Q, when each $H_1, \dots, H_p$, $B_1, \dots, B_q$, and $b_1, \dots, b_q$ is simple. Combining adjacent x's and y's in Q, and extending the families $\{H_1, \dots, H_p\}$ and $\{B_1, \dots, B_q\}$ to include products and Hermitian conjugates of these matrices as necessary, we may assume that Q is an alternating word in x_i's and y_i's. If Q begins with x_i or ends with x_j, let us use $\tau^{\mathcal{H}}(H_i a H_j) = H_i \tau^{\mathcal{H}}(a) H_j$ and replace v by $H_i^* v$ and w by $H_j w$. Then the result follows from Lemma 4.5. □

Proof of Lemma 4.5. We induct on m. The result is clear for $m = 0$, as $\tau^{\mathcal{H}}(1) = 1$ and the left side of (4.5) is simply $v^*w - v^*w$. Suppose by induction that the lemma holds up to $m-1$, and consider the case of m. Introduce the centered elements

$$\mathring{H}_i = H_i - \tau^{\mathcal{D}}(H_i), \qquad \mathring{B}_j = B_j - \tau^{\mathcal{D}}(b_j), \qquad \mathring{b}_j = b_j - \tau^{\mathcal{D}}(b_j).$$

(Note that here, we first center B_j by $\tau^{\mathcal{D}}(b_j)$, not a normalized trace of B_j.) On the left side of (4.5), let us write $H_{i_r} = \mathring{H}_{i_r} + \tau^{\mathcal{D}}(H_{i_r})$ for each i_r, and similarly for each B_{j_r} and b_{j_r}. Expanding the resulting product, we obtain that the left side of (4.5) is equal to

$$v^* \mathring{B}_{j_1} \mathring{H}_{i_1} \dots \mathring{B}_{j_{m-1}} \mathring{H}_{i_{m-1}} \mathring{B}_{j_m} w - v^* \tau^{\mathcal{H}}(\mathring{b}_{j_1} \mathring{H}_{i_1} \dots \mathring{b}_{j_{m-1}} \mathring{H}_{i_{m-1}} \mathring{b}_{j_m}) w \tag{4.6}$$

plus a (constant) number of remainder terms which include at least one factor $\tau^{\mathcal{D}}(H_i)$ or $\tau^{\mathcal{D}}(b_j)$. Since H_i is simple, we have either $\tau^{\mathcal{D}}(H_i) = 0$ or $\tau^{\mathcal{D}}(H_i) = z(H_i) \cdot P_{r_i}$ for some $r_i \in \{1, \dots, k\}$ and for $z(H_i) = \tau(H_i)/\tau(P_{r_i}) \in \mathbb{C}$, and similarly for $\tau^{\mathcal{D}}(b_j)$. Then, absorbing P_{r_i} into the adjacent factor and applying the arguments of the proof of Theorem 4.4 above, each such remainder term may be written as a sum of differences of the form (4.5) for a value $m' \le m-1$, multiplied by an N-dependent coefficient z_N which is a product of a subset of the coefficients $z(H_1), \dots, z(H_p), z(b_1), \dots, z(b_q)$. Since $\|\tau^{\mathcal{D}}(H_i)\| \le \|H_i\| \le C$ and similarly for b_j, we have that $|z_N| \le C$ for a constant $C > 0$ and all N. Then the remainder terms converge to 0 by the inductive hypothesis.

It remains to show that the difference (4.6) converges to 0. We claim that

$$\tau^{\mathcal{H}}(\mathring{b}_{j_1} \mathring{H}_{i_1} \dots \mathring{b}_{j_{m-1}} \mathring{H}_{i_{m-1}} \mathring{b}_{j_m}) = 0. \tag{4.7}$$

Indeed, letting $\mathrm{NC}(m)$ be the set of non-crossing partitions of $\{1, \dots, m\}$ and introducing the $\mathcal{H}$-valued non-crossing cumulants $\kappa_\pi^{\mathcal{H}}$, we have

$$\tau^{\mathcal{H}}(\mathring{b}_{j_1} \mathring{H}_{i_1} \dots \mathring{b}_{j_{m-1}} \mathring{H}_{i_{m-1}} \mathring{b}_{j_m}) = \sum_{\pi \in \mathrm{NC}(m)} \kappa_\pi^{\mathcal{H}}(\mathring{b}_{j_1} \mathring{H}_{i_1}, \ \dots, \mathring{b}_{j_{m-1}} \mathring{H}_{i_{m-1}}, \ \mathring{b}_{j_m}).$$

Each partition π has an element which is an interval $\{r, \dots, r+\ell-1\}$ of consecutive indices, for some $\ell \ge 1$. Letting $\tau^{\mathcal{D}}$ be the τ-invariant projection onto $\mathcal{D}$, we apply [NSS02, Theorem 3.5] and freeness of $\mathcal{H}$ and $\mathcal{B}$ over $\mathcal{D}$ to obtain

$$\kappa_\ell^{\mathcal{H}}(b_1 H_1, \dots, b_{\ell-1} H_{\ell-1}, b_\ell) = \kappa_\ell^{\mathcal{D}}(b_1 \tau^{\mathcal{D}}(H_1), \dots, b_{\ell-1} \tau^{\mathcal{D}}(H_{\ell-1}), b_\ell) = 0$$

for any elements $b_1, \dots, b_\ell \in \mathcal{B}$ and $H_1, \dots, H_{\ell-1} \in \mathcal{H}$ which are zero-centered with respect to $\tau^{\mathcal{D}}$. (In the case $\ell = 1$, the latter equality holds because $\kappa_1^{\mathcal{D}}(b_1) = \tau^{\mathcal{D}}(b_1) = 0$.) Applying this to the cumulant $\kappa_\ell^{\mathcal{H}}$ of the terms corresponding to this interval $\{r, \dots, r+\ell-1\}$ of π, we obtain $\kappa_\pi^{\mathcal{H}}(\mathring{b}_{j_1} \mathring{H}_{i_1}, \ \dots, \mathring{b}_{j_{m-1}} \mathring{H}_{i_{m-1}}, \ \mathring{b}_{j_m}) = 0$ for each $\pi \in \mathrm{NC}(m)$, and hence (4.7).

Thus, to show that (4.6) converges to 0, we must show that correspondingly

$$v^* \mathring{B}_{j_1} \mathring{H}_{i_1} \dots \mathring{B}_{j_{m-1}} \mathring{H}_{i_{m-1}} \mathring{B}_{j_m} w \to 0. \tag{4.8}$$

Since H_i and B_j are simple, some (r_i, s_i) block of each $\mathring{H}_i$ is non-zero and the remaining blocks are 0, and some (t_j, u_j) block of each $\mathring{B}_j$ is non-zero and the remaining blocks are 0. We may suppose $u_{j_1} = r_{i_1}$, $s_{i_1} = t_{j_2}$, $u_{j_2} = r_{i_2}$, etc., for otherwise the left side of (4.8) is automatically 0. Denote by

$$\check{H}_i \in \mathbb{C}^{N_{r_i} \times N_{s_i}}$$

the non-zero block of $\mathring{H}_i$. If $r_i \neq s_i$, then $\check{H}_i$ is just the corresponding block $(H_i)_{r_i s_i}$ of H_i. If $r_i = s_i$, then by the fact that τ coincides with $N^{-1}\operatorname{Tr}$ on $\mathcal{A}_1$, $\check{H}_i = (H_i)_{r_i r_i} - N_{r_i}^{-1}\operatorname{Tr} H_i$ is the centered version of this block. Define also

$$\check{B}_j \in \mathbb{C}^{N_{t_j} \times N_{u_j}}$$

to be the non-zero block of $\mathring{B}_j$ if $t_j \neq u_j$, or $\check{B}_j = (B_j)_{t_j t_j} - N_{t_j}^{-1}\operatorname{Tr} B_j$ if $t_j = u_j$. In the latter case, note that $\check{B}_j$ differs from the nonzero block of $\mathring{B}_j$ by the quantity

$$\left(\frac{N}{N_{t_j}}\tau(B_j) - N_{t_j}^{-1}\operatorname{Tr} B_j\right)\operatorname{Id}_{N_{t_j}} \to 0, \tag{4.9}$$

where the convergence is in operator norm as $N \to \infty$ by Assumption 4.1. Finally, define $\check{v} \in \mathbb{C}^{r_{j_1}}$ to be the r_{j_1} block of v, and $\check{w} \in \mathbb{C}^{s_{j_m}}$ to be the s_{j_m} block of w. Then

$$\left|v^* \mathring{B}_{j_1} \mathring{H}_{i_1} \dots \mathring{B}_{j_{m-1}} \mathring{H}_{i_{m-1}} \mathring{B}_{j_m} w - \check{v}^* \check{B}_{j_1} \check{H}_{i_1} \dots \check{B}_{j_{m-1}} \check{H}_{i_{m-1}} \check{B}_{j_m} \check{w}\right| \to 0,$$

almost surely as $N \to \infty$, by the observation (4.9) and the operator norm bound on each H_i and B_j. So it suffices to show

$$\check{v}^* \check{B}_{j_1} \check{H}_{i_1} \dots \check{B}_{j_{m-1}} \check{H}_{i_{m-1}} \check{B}_{j_m} \check{w} \to 0.$$

Let us introduce a random orthogonal matrix

$$O = \operatorname{diag}(O_1, \dots, O_k) \in \mathcal{O}$$

where each $O_r \in \mathbb{R}^{N_r \times N_r}$ is independently Haar-distributed on the orthogonal group and also independent of $B_1, \dots, B_q$. By the assumed conjugation invariance of $(B_1, \dots, B_q)$, we have the equality in law

$$(\check{B}_1, \dots, \check{B}_q) \overset{L}{=} (O_{t_1} \check{B}_1 O_{u_1}^{-1}, \dots, O_{t_q} \check{B}_q O_{u_q}^{-1}),$$

and thus we may equivalently show (almost surely as $N \to \infty$)

$$\check{v}^* O_{t_{j_1}} \check{B}_{j_1} O_{u_{j_1}}^{-1} \check{H}_{i_1} \dots O_{t_{j_m}} \check{B}_{j_m} O_{u_{j_m}}^{-1} \check{w} \to 0. \tag{4.10}$$

We then condition on $\check{B}_1, \dots, \check{B}_q$, and write $\mathbb{E}$ for the expectation over $O_1, \dots, O_k$. Defining

$$\mathcal{E} = \mathbb{E}\Big[|\check{v}^* O_{t_{j_1}} \check{B}_{j_1} O_{u_{j_1}}^{-1} \check{H}_{i_1} \dots O_{t_{j_m}} \check{B}_{j_m} O_{u_{j_m}}^{-1} \check{w}|^4\Big],$$

we observe that this may be written in the form

$$\mathcal{E} = \mathbb{E}[\operatorname{Tr} O_{r_1}^{e_1} D_1 O_{r_2}^{e_2} D_2 \dots O_{r_{8m}}^{e_{8m}} D_{8m}]$$

where

- Each $r_i \in \{1, \dots, k\}$ and each $e_i \in \{-1, 1\}$.
- Each D_i is one of $\check{H}_1, \dots, \check{H}_p$, $\check{B}_1, \dots, \check{B}_q$, $\check{w}\check{v}^*$, $\check{w}\check{v}^{\mathsf{T}}$ or their Hermitian conjugates.
- If $r_i = r_{i+1}$ and D_i is not of the form $\check{w}\check{v}^*$, $\check{w}\check{v}^{\mathsf{T}}$ or their conjugates, then the centering of $\check{H}$ and $\check{B}$ implies $\operatorname{Tr} D_i = 0$.
- At least four of the matrices $D_1, \dots, D_{8m}$ are of rank 1.

Then Lemma 4.6 below implies (conditional on $\check{B}_1, \dots, \check{B}_q$ for all N, and on the event of probability 1 where $\|\check{B}_1\|, \dots, \|\check{B}_q\| < C'$ for a constant $C' > 0$ and all large N) that $\mathcal{E} \leq CN^{-2}$. Then (4.10) holds almost surely as $N \to \infty$ by Markov's inequality and Borel-Cantelli, as desired. □

Lemma 4.6. *Fix constants $B, C, c > 0$ and suppose $c < N_r/N < C$ for each $r = 1, \dots, k$. Let $O_1, \dots, O_k$ be independent matrices, with each $O_r \in \mathbb{R}^{N_r \times N_r}$ Haar-distributed on the orthogonal group.*

Fix $M \geq 1$, $r_1, \dots, r_M \in \{1, \dots, k\}$, $e_1, \dots, e_M \in \{-1, 1\}$, and cyclically identify $r_{M+1} \equiv r_1$. For each $m = 1, \dots, M$, let $D_m \in \mathbb{C}^{N_{r_m} \times N_{r_{m+1}}}$ be a deterministic matrix with $\|D_m\| < B$. For each m, suppose at least one of the following holds:

- *$r_m \neq r_{m+1}$, or*
- *D_m is of rank 1, or*
- *$r_m = r_{m+1}$ and $\operatorname{Tr} D_m = 0$.*

Finally, suppose that at least K of $D_1, \dots, D_M$ have rank 1. Then for a constant $C' \equiv C'(k, K, M, B) > 0$,

$$\mathbb{E}[\operatorname{Tr} O_{r_1}^{e_1} D_1 O_{r_2}^{e_2} D_2 \dots O_{r_M}^{e_M} D_M] \leq C' N^{-K/2}.$$

Proof. The proof of this lemma is similar to that of [FJ16, Lemma B.2], which established a version of this result for $K = 0$. We extend the combinatorial argument here to handle the case of general K. To ease subscript notation, we write $v[i]$ and $A[i,j]$ for entry i of v and entry (i,j) of A. We denote by $C > 0$ a constant which may depend on k, K, M, B and change from instance to instance.

We may write

$$\mathcal{E} \equiv \mathbb{E}[\operatorname{Tr} O_{r_1}^{e_1} D_1 O_{r_2}^{e_2} D_2 \dots O_{r_M}^{e_M} D_M] = \sum_{\mathbf{i},\mathbf{j}} D(\mathbf{i},\mathbf{j}) \mathbb{E}[V(\mathbf{i},\mathbf{j})],$$

where the sum is over all tuples $(\mathbf{i},\mathbf{j}) = (i_1, \dots, i_M, j_1, \dots, j_M)$ satisfying

$$1 \le i_k, j_k \le N_{r_k}$$

and where

$$V(\mathbf{i},\mathbf{j}) = \prod_{m=1}^{M} O_{r_m}^{e_m}[i_m, j_m], \qquad D(\mathbf{i},\mathbf{j}) = \prod_{m=1}^{M} D_m[j_m, i_{m+1}]$$

with the cyclic identification $i_{M+1} \equiv i_1$. Define the set partition

$$\bigsqcup_{r=1}^{k} \mathcal{I}(r) = \{1, \dots, M\}$$

by $\mathcal{I}(r) = \{m : r_m = r\}$. Consider now set partitions of the set $\{1, \dots, M\} \sqcup \{1, \dots, M\}$ of cardinality $2M$, where we denote elements of the first copy of $\{1, \dots, M\}$ with a subscript i and the second with a subscript j. A set in this partition can have elements of either or both copies of $\{1, \dots, M\}$; for example, $\{1_i, 2_j\}$ or $\{2_j, 3_j\}$ might be sets in the set partition. We say that $\mathbf{i},\mathbf{j}$ induces $\mathcal{Q}$, denoted $\mathbf{i},\mathbf{j} \mid \mathcal{Q}$, if

$$\mathcal{Q} = \bigsqcup_{r=1}^{k} \bigsqcup_{s=1}^{N_r} (\mathcal{Q}^1(r,s) \sqcup \mathcal{Q}^2(r,s)),$$

for

$$\mathcal{Q}^1(r,s) = \{m_i : m \in \mathcal{I}(r), i_m = s, e_m = 1\} \cup \{m_j : m \in \mathcal{I}(r), j_m = s, e_m = -1\}$$
$$\mathcal{Q}^2(r,s) = \{m_i : m \in \mathcal{I}(r), i_m = s, e_m = -1\} \cup \{m_j : m \in \mathcal{I}(r), j_m = s, e_m = 1\}.$$

Denote $\mathcal{Q}(r) := \bigsqcup_{s=1}^{N_r} (\mathcal{Q}^1(r,s) \sqcup \mathcal{Q}^2(r,s))$, and let $|\mathcal{Q}|$ be the total number of non-empty sets in $\mathcal{Q}$.

Notice that the quantity

$$\mathbb{E}[V(\mathbf{i},\mathbf{j})] \equiv E(\mathcal{Q})$$

depends on $(\mathbf{i},\mathbf{j})$ only via its induced partition $\mathcal{Q}$. By [FJ16, Lemma B.3(a)] we have $|E(\mathcal{Q})| \le CN^{-M/2}$ for any partition $\mathcal{Q}$. Thus we find

$$\mathcal{E} \le CN^{-M/2} \sum_{\mathcal{Q}: E(\mathcal{Q}) \ne 0} |D(\mathcal{Q})|, \qquad D(\mathcal{Q}) \equiv \sum_{\mathbf{i},\mathbf{j} | \mathcal{Q}} D(\mathbf{i},\mathbf{j}), \tag{4.11}$$

so our main task is to bound $|D(\mathcal{Q})|$ when $E(\mathcal{Q}) \ne 0$. By [FJ16, Lemma B.3(b)], if $\mathbf{i},\mathbf{j} \mid \mathcal{Q}$ and $E(\mathcal{Q}) \ne 0$, then for each $r \in \{1, \dots, k\}$ and each $s \in \{1, \dots, N_r\}$, the cardinality of $|\mathcal{Q}^1(r,s)|$ and $|\mathcal{Q}^2(r,s)|$ must be even. That is, each set $S \in \mathcal{Q}$ has even cardinality. To motivate the combinatorial idea, note that the bound $|D_m[j_m, i_{m+1}]| \le B$ implies that $D(\mathbf{i},\mathbf{j}) \le B^M$ for all $(\mathbf{i},\mathbf{j})$, while

$$\#\{(\mathbf{i},\mathbf{j}) : \mathbf{i},\mathbf{j} \mid \mathcal{Q}\} \le CN^{|\mathcal{Q}|},$$

since for any fixed $\mathcal{Q}$ choosing $\mathbf{i},\mathbf{j}$ which induce $\mathcal{Q}$ involves choosing for each set in $\mathcal{Q}$ a distinct index from $\{1, \dots, N_r\}$ for some r. Together, these yield the naive bound $|D(\mathcal{Q})| \le CN^{|\mathcal{Q}|}$. Since each set in $\mathcal{Q}$ has cardinality at least 2, and the sum of all cardinalities is $2M$, we have $|\mathcal{Q}| \le M$. Combining with (4.11) would yield

$$\mathcal{E} \le CN^{-M/2} \cdot N^M,$$

but the exponent is too large in M and does not depend on the number of rank 1 matrices K.

This motivates the definitions of the following counts associated to $\mathcal{Q}$. For $m \in \{1, \dots, M\}$, call the index m_i single if D_{m-1} is of rank 1 and the index m_j single if D_m is of rank 1—that is, an index is single if it corresponds to some rank 1 matrix in the product $D(\mathbf{i},\mathbf{j})$. For a fixed set partition $\mathcal{Q}$, define the following quantities.

- T_0: number of sets in $\mathcal{Q}$ of cardinality 2, which contain no single indices.
- T_1: number of sets in $\mathcal{Q}$ of cardinality 2, which contain 1 or 2 single indices.
- R_0: number of sets in $\mathcal{Q}$ of cardinality ≥ 4, which contain no single indices.
- R_1: number of sets in $\mathcal{Q}$ of cardinality ≥ 4, which contain (exactly) 1 single index.

We establish the following claim by induction on $T_0 + T_1$.

Inductive claim: For any $M \geq 1$, any $r_1, \dots, r_M, e_1, \dots, e_M, D_1, \dots, D_M$ which satisfy the conditions of the lemma, and any such partition $\mathcal{Q}$ of $\{1, \dots, M\} \sqcup \{1, \dots, M\}$ with T_0, T_1, R_0, R_1 as defined above,

$$|D(\mathcal{Q})| \leq C_0 N^{R_0 + T_0/2 + R_1/2} \tag{4.12}$$

for a constant $C_0 \equiv C_0(k, M, T_0, T_1, R_0, R_1, B) > 0$.

Assuming that this claim holds, note that the number of non-single indices is $2(M - K)$, where K is the number of rank 1 matrices. Then $2(M - K) \geq 4R_0 + 2T_0 + 3R_1$. Dividing this by 4 gives the improved bound

$$|D(\mathcal{Q})| \leq C_0 N^{(M-K)/2}.$$

Combining with (4.11) yields $\mathcal{E} \leq CN^{-K/2}$, as desired.

To establish (4.12), we induct on the total number of elements of $\mathcal{Q}$ of cardinality 2, which is $T_0 + T_1$. For the base case $T_0 + T_1 = 0$, let us assume for notational convenience that $D_1, \dots, D_K$ are of rank 1. For $m = 1, \dots, K$, we write $D_m = v_m w_m^*$ for bounded length vectors v_m and w_m, and apply $|D_m[i, j]| \leq B$ for $m = K + 1, \dots, M$. This gives

$$|D(\mathcal{Q})| \leq C \sum_{\mathbf{i},\mathbf{j}|\mathcal{Q}} |v_1[j_1] w_1[i_2] \cdots v_K[j_K] w_K[i_{K+1}]|. \tag{4.13}$$

Let R_2 be the number of elements of $\mathcal{Q}$ containing two or more single indices. Since $\mathcal{Q}$ has no elements of cardinality 2, all elements of $\mathcal{Q}$ are counted by R_0, R_1, or R_2. We now view the sum in (4.13) as a product of sums over distinct indices for the elements of $\mathcal{Q}$ counted by R_0, R_1, R_2. We bound the sum over distinct indices counted by R_0 simply by CN^{R_0}. For the sum over distinct indices counted by R_1, note by Cauchy-Schwartz that

$$\sum_i |u[i]| \leq \sqrt{\|u\|} \sqrt{N},$$

yielding a combined bound of $CN^{R_1/2}$ for these indices because $\|u\|$ is bounded for the relevant vectors. For distinct indices counted by R_2, we apply a bound of the form

$$\sum_i |u_1[i] \dots u_m[i]| \leq C \sum_i |u_1[i] u_2[i]| \leq C \|u_1\| \cdot \|u_2\|$$

for any $m \geq 2$ and any bounded vectors $u_1, \dots, u_m$, yielding a constant bound for the combined sum over such indices. Thus, we get

$$|D(\mathcal{Q})| \leq CN^{R_0 + R_1/2},$$

which concludes the proof of (4.12) in this base case.

Assume inductively that (4.12) holds for $T_0 + T_1 \leq t - 1$, and consider now $T_0 + T_1 = t \geq 1$. Then there is some set $S \in \mathcal{Q}$ with cardinality $|S| = 2$. We consider three cases.

Case 1: $S = \{m_j, (m + 1)_i\}$, and D_m is *not* of rank 1. (So S is counted by T_0.) Suppose for notational convenience that $S = \{1_j, 2_i\}$. This implies in particular that $r_1 = r_2$ and D_1 is square. Then the assumption of the lemma implies

$$\operatorname{Tr} D_1 = 0.$$

Denote by $\sum_{\mathbf{i},\mathbf{j}|\mathcal{Q}\setminus S}$ the sum over indices in the tuple $(\mathbf{i}, \mathbf{j})$ excluding j_1 and i_2 which induce $\mathcal{Q} \setminus S$, and by $\sum_{j \notin \mathcal{Q}(r_1)\setminus S}$ the remaining sum over the value of $j_1 \equiv i_2$, restricted to be distinct from the $|\mathcal{Q}(r_1)| - 1$ preceding values in $\{1, \dots, N_{r_1}\}$ assumed by sets in $\mathcal{Q}(r_1) \setminus S$. Then

$$D(\mathcal{Q}) = \sum_{\mathbf{i},\mathbf{j}|\mathcal{Q}\setminus S} \prod_{m=2}^{M} D_m[j_m, i_{m+1}] \cdot \sum_{j \notin \mathcal{Q}(r_1)\setminus S} D_1[j, j].$$

Let Π be the set of new partitions $\mathcal{Q}'$ which merge $S = \{1_j, 2_i\}$ with some other set in $\mathcal{Q}(r_1) \setminus S$. Then applying $\operatorname{Tr} D_1 = 0$ yields

$$D(\mathcal{Q}) = -\sum_{\mathcal{Q}' \in \Pi} \sum_{\mathbf{i},\mathbf{j}|\mathcal{Q}'} \prod_{m=1}^{M} D_m[j_m, i_{m+1}],$$

and hence

$$|D(\mathcal{Q})| \le \sum_{\mathcal{Q}' \in \Pi} |D(\mathcal{Q}')|.$$

As D_m is not of rank 1, the indices $1_j, 2_i$ are not single. If $\{1_j, 2_i\}$ was merged into a set in $\mathcal{Q}$ of cardinality ≥ 4, then $\mathcal{Q}'$ has the counts $(T_0 - 1, T_1, R_0, R_1)$. If $\{1_j, 2_i\}$ was merged into a set in $\mathcal{Q}$ counted by T_1, then $\mathcal{Q}'$ has either the counts $(T_0 - 1, T_1 - 1, R_0, R_1)$ or $(T_0 - 1, T_1 - 1, R_0, R_1 + 1)$. If $\{1_j, 2_i\}$ was merged into another set in $\mathcal{Q}$ counted by T_0, then $\mathcal{Q}'$ has the counts $(T_0 - 2, T_1, R_0 + 1, R_1)$. In all cases, $T_0 + T_1$ has reduced by at least 1, and the exponent $R_0 + T_0/2 + R_1/2$ in (4.12) has not increased. Then applying the inductive hypothesis for each $\mathcal{Q}'$ and noting that the cardinality of Π is a constant independent of N, we get (4.12) for $\mathcal{Q}$.

Case 2: $S = \{m_j, (m+1)_i\}$, and D_m *is* of rank 1. (So S is counted by T_1.) Suppose for notational convenience $S = \{1_j, 2_i\}$. Then with the same notation as defined in Case 1, we get

$$|D(\mathcal{Q})| \le |\operatorname{Tr} D_1| \cdot \left| \sum_{\mathbf{i},\mathbf{j}|\mathcal{Q}\setminus S} \prod_{m=2}^{M} D_m[j_m, i_{m+1}] \right| + \sum_{\mathcal{Q}' \in \Pi} \left| \sum_{\mathbf{i},\mathbf{j}|\mathcal{Q}'} \prod_{m=1}^{M} D_m[j_m, i_{m+1}] \right|,$$

where the first term arises because we no longer have $\operatorname{Tr} D_1 = 0$. (If $M = 1$, the first term is understood to just be $|\operatorname{Tr} D_1|$.) Note that $|\operatorname{Tr} D_1| \le C$, as D_1 has bounded operator norm and is of rank 1. The partition $\mathcal{Q} \setminus S$ in the first term must have the counts $(T_0, T_1 - 1, R_0, R_1)$, and we may apply the inductive hypothesis to this term. For each $\mathcal{Q}'$ in the second term, the argument is a bit different from Case 1 as $1_j, 2_i$ are single. If $\{1_j, 2_i\}$ was merged into a set in $\mathcal{Q}$ counted by T_0, T_1, R_0, R_1, or none of these four, then $\mathcal{Q}'$ has the counts $(T_0 - 1, T_1 - 1, R_0, R_1)$, $(T_0, T_1 - 2, R_0, R_1)$, $(T_0, T_1 - 1, R_0 - 1, R_1)$, $(T_0, T_1 - 1, R_0, R_1 - 1)$, or $(T_0, T_1 - 1, R_0, R_1)$ respectively. Applying the inductive hypothesis in all cases, we get (4.12) for $\mathcal{Q}$.

Case 3: The two indices in S do not index the same matrix D_m. Suppose for notational convenience $S = \{2_i, 2_j\}$, so that they index D_1 and D_2; other cases are analogous. Then with similar notation as in Case 1, we have

$$D(\mathcal{Q}) = \sum_{\mathbf{i},\mathbf{j}|\mathcal{Q}\setminus S} \prod_{m=3}^{M} D_m[j_m, i_{m+1}] \cdot \sum_{i \notin \mathcal{Q}(r_2)\setminus S} D_1[j_1, i] D_2[i, i_3].$$

Let us introduce the matrix $\tilde{D} = D_1 D_2$. Then applying the triangle inequality as in Cases 1 and 2,

$$|D(\mathcal{Q})| \le \left| \sum_{\mathbf{i},\mathbf{j}|\mathcal{Q}\setminus S} \tilde{D}[j_1, i_3] \prod_{m=3}^{M} D_m[j_m, i_{m+1}] \right| + \sum_{\mathcal{Q}' \in \Pi} \left| \prod_{m=1}^{M} D_m[j_m, i_{m+1}] \right|,$$

where Π is the set of partitions $\mathcal{Q}'$ which merge $\{2_i, 2_j\}$ with another set in of $\mathcal{Q}(r_2) \setminus S$. (The product in the first term is understood to be 1 if $M = 2$.)

For the first term involving $\mathcal{Q} \setminus S$, note that if $\tilde{D}$ is not of rank 1, then both D_1 and D_2 are also not of rank 1. So $2_i, 2_j, 1_j, 3_i$ were not single in $\mathcal{Q}$, and $1_j, 3_i$ remain non-single in $\mathcal{Q} \setminus S$ (with respect to $\tilde{D}, D_3, \ldots, D_M$). Then $\mathcal{Q} \setminus S$ must have the counts $(T_0 - 1, T_1, R_0, R_1)$. If $\tilde{D}$ is of rank 1, then the removal of $\{2_i, 2_j\}$ reduces either T_0 or T_1 by 1, but it is possible that 1_j and/or 3_i may have been converted from a non-single index in $\mathcal{Q}$ to a single index in $\mathcal{Q} \setminus S$. One such conversion may induce the count mapping $(T_0, T_1) \mapsto (T_0, T_1 - 1)$, $(T_0, T_1) \mapsto (T_0 - 1, T_1 + 1)$, $(R_0, R_1) \mapsto (R_0, R_1 - 1)$, or $(R_0, R_1) \mapsto (R_0 - 1, R_1 + 1)$. Note that each of these mappings does not increase $T_0 + T_1$, nor increase the exponent $R_0 + T_0/2 + R_1/2$ of N in (4.12). Then we may apply the induction hypothesis in every case to obtain $|D(\mathcal{Q} \setminus S)| \le C N^{R_0 + T_0/2 + R_1/2}$ for the first term.

For each $\mathcal{Q}' \in \Pi$ of the second term, we perform some casework, depending on whether $2_i, 2_j$ are both non-single (so D_1 and D_2 both have rank more than 1), and also whether $\{2_i, 2_j\}$ was merged into a set in $\mathcal{Q}$ counted by T_0, T_1, R_0, R_1 or none of these four. The possible resulting counts for $\mathcal{Q}'$ are summarized in Table 4.1. In each setting, $T_0 + T_1$ has reduced by at least 1, the exponent $R_0 + T_0/2 + R_1/2$ has not increased, and we may thus apply the induction hypothesis for $\mathcal{Q}'$ to obtain (4.12) for $\mathcal{Q}$.

Merged into	$2_j, 2_i$ not single	one or both of $2_j, 2_i$ single
T_0	T_0-2, T_1, R_0+1, R_1	T_0-1, T_1-1, R_0, R_1 or T_0-1, T_1-1, R_0, R_1+1
T_1	T_0-1, T_1-1, R_0, R_1 or T_0-1, T_1-1, R_0, R_1+1	T_0, T_1-2, R_0, R_1
R_0	T_0-1, T_1, R_0, R_1	T_0, T_1-1, R_0-1, R_1 or T_0, T_1-1, R_0-1, R_1+1
R_1	T_0-1, T_1, R_0, R_1	T_0, T_1-1, R_0, R_1-1
None of above	T_0-1, T_1, R_0, R_1	T_0, T_1-1, R_0, R_1

TABLE 4.1. Possible counts for $\mathcal{Q}'$

This establishes that (4.12) holds when $T_0+T_1=t$, in all three of the above Cases. This completes the induction and the proof of the lemma. □

4.3. **Convergence for resolvent.** Finally, we use Theorem 4.4 to complete the proofs of Theorem 4.2 and Corollary 4.3. This will depend on the following lemma, which allows us to work with a series expansion of the Stieltjes transform.

Lemma 4.7. *Let $C>0$ be such that $\|W\|\le C$ and $\|w\|\le C$ for large N, and suppose that f_N is an analytic function on $\mathbb{C}\setminus\operatorname{spec}(W)$ and f an analytic function on $\mathbb{C}\setminus\operatorname{spec}(w)$ such that almost surely as $N\to\infty$, we have $f_N - f\to 0$ uniformly on $\mathbb{D}'=\{z\in\mathbb{C}:|z|>2C\}$. Then for any fixed constant $\delta>0$, almost surely, $f_N-f\to 0$ uniformly on $\mathbb{D}_N=\{z\in\mathbb{C}:\operatorname{dist}(z,\operatorname{spec}(w))\ge\delta$ and $\operatorname{dist}(z,\operatorname{spec}(W))\ge\delta\}$.*

Proof. Let Ω_0 be the event of probability 1 where $\operatorname{spec}(W)$ (and also $\operatorname{spec}(w)$) are uniformly bounded in $[-C,C]$ for all large N, and

$$\lim_{N\to\infty}\sup_{z\in\mathbb{D}'}|f_N(z)-f(z)|=0.$$

Suppose by contradiction that for some $\omega\in\Omega_0$ and $\varepsilon>0$, we have

$$\limsup_{N\to\infty}\sup_{z\in\mathbb{D}_N}|f_N(z)-f(z)|>\varepsilon. \tag{4.14}$$

Then there is a subsequence $\{N_k\}_{k=1}^\infty$ and points $z_{N_k}\in\mathbb{D}_{N_k}$ for which $|f_{N_k}(z_{N_k})-f(z_{N_k})|>\varepsilon$ for all k. Since $\operatorname{spec}(W)$ and $\operatorname{spec}(w)$ are uniformly bounded compact subsets of $\mathbb{R}$, by sequential compactness under Hausdorff distance, there must be a further subsequence of $\{N_k\}_{k=1}^\infty$ along which these sets converge in Hausdorff distance to fixed limits $S_1\equiv S_1(\omega)$ and $S_2\equiv S_2(\omega)$. Define $\mathbb{D}_\infty(\omega)=\{z\in\mathbb{C}:\operatorname{dist}(z,S_1)\ge\delta/2,\operatorname{dist}(z,S_2)\ge\delta/2\}$. Then $\mathbb{D}_\infty(\omega)$ is a fixed (N-independent) connected domain of $\mathbb{C}$. As $f_N(z)-f(z)$ is analytic on $\mathbb{D}_\infty(\omega)$ for all large N, we then have

$$\lim_{N\to\infty}\sup_{z\in\mathbb{D}_\infty(\omega)}|f_N(z)-f(z)|=0,$$

by the convergence over $z\in\mathbb{D}'$. This implies $z_{N_k}\notin\mathbb{D}_\infty(\omega)$ for all large k. But then

$$\limsup_{k\to\infty}\min(\operatorname{dist}(z_{N_k},S_1),\operatorname{dist}(z_{N_k},S_2))\le\delta/2,$$

which implies by the definition of Hausdorff distance that

$$\limsup_{k\to\infty}\min(\operatorname{dist}(z_{N_k},\operatorname{spec}(w)),\operatorname{dist}(z_{N_k},\operatorname{spec}(W)))\le\delta/2,$$

contradicting that $z_{N_k}\in\mathbb{D}_{N_k}$. Thus (4.14) cannot hold for any $\omega\in\Omega_0$. □

Proof of Theorem 4.2. The given assumptions imply that there is a constant $C>0$ such that $\|W\|\le C$ and $\|w\|\le C$ almost surely for all large N. Let $\mathbb{D}'=\{z\in\mathbb{C}:|z|>2C\}$. Fix $\varepsilon>0$. Applying the contractive property $\|\tau^{\mathcal{H}}(a)\|\le\|a\|$ of conditional expectations, there is $K>0$ such that

$$\sup_{z\in\mathbb{D}'}\left\|\sum_{k=K+1}^\infty z^{-(k+1)}W^k\right\|<\varepsilon,\qquad\sup_{z\in\mathbb{D}'}\left\|\sum_{k=K+1}^\infty z^{-(k+1)}\tau^{\mathcal{H}}(w^k)\right\|<\varepsilon$$

for all large N. For each $k \in \{0, \dots, K\}$, Theorem 4.4 implies $u^* W^k v - u^* \tau^{\mathcal{H}}(w^k) v \to 0$ almost surely. Then applying the series expansions for $(w - z)^{-1}$ and $(W - z\,\mathrm{Id})^{-1}$, convergent for $|z| > 2C$, we get

$$\limsup_{N\to\infty} \sup_{z\in\mathbb{D}'} |u^*(W - z\,\mathrm{Id})^{-1} v - u^* R_0(z) v| < 2\varepsilon.$$

As $\varepsilon > 0$ is arbitrary, we obtain almost surely

$$\lim_{N\to\infty} \sup_{z\in\mathbb{D}'} |u^*(W - z\,\mathrm{Id})^{-1} v - u^* R_0(z) v| = 0. \tag{4.15}$$

Applying Lemma 4.7 for $f_N(z) = u^*(W - z\,\mathrm{Id})^{-1} v$ and $f(z) = u^* R_0(z) v$ concludes the proof. □

Proof of Corollary 4.3. Let $W' = W + P_2 + \ldots + P_k$ and $w' = w + P_2 + \ldots + P_k$. Note that W', w' define the same submatrices $W_{11} \in \mathbb{C}^{N_1 \times N_1}$ and $(R_0(z))_{11} \in \mathbb{C}^{N_1 \times N_1}$, the latter because

$$P_1 \tau^{\mathcal{H}}((w' - z)^{-1}) P_1 = \tau^{\mathcal{H}}(P_1 (w' - z)^{-1} P_1) = \tau^{\mathcal{H}}(P_1 (w - z)^{-1} P_1) = P_1 \tau^{\mathcal{H}}((w - z)^{-1}) P_1.$$

On the other hand, for $k \geq 2$, their spectra satisfy

$$\mathrm{spec}(W) = \mathrm{spec}(W_{11}) \cup \{0\}, \quad \mathrm{spec}(w) = \mathrm{spec}(w_{11}) \cup \{0\},$$

$$\mathrm{spec}(W') = \mathrm{spec}(W_{11}) \cup \{1\}, \quad \mathrm{spec}(w) = \mathrm{spec}(w_{11}) \cup \{1\}.$$

Then for any $\delta \leq 1/2$, setting $\mathbb{D}$ and $\mathbb{D}'$ as the sets (4.2) with (W, w) and (W', w'), we have

$$\mathbb{D}_1 = \mathbb{D} \cup \mathbb{D}'.$$

Then the result follows from applying Theorem 4.2 with $u = (u_1, 0, \dots, 0)$ and $v = (v_1, 0, \dots, 0)$, for both (W, w) and (W', w'). □

5. Analysis of the mixed effects model

In this section, we prove the results stated in Section 2.3 pertaining to the mixed effects model.

5.1. **Preliminary results.** First, we prove Theorem 2.4, which guarantees that no bulk eigenvalues separate from the support.

Proof of Theorem 2.4. We consider the block decomposition (2.10) in $\mathbb{C}^{N\times N}$, the orthogonal projections $P_0, \dots, P_{2k}$, and the embedded matrices $\tilde{F}_{rs}, \tilde{G}_r, \tilde{H}_r \in \mathbb{C}^{N\times N}$. Then the only non-zero block of the matrix

$$\tilde{W} = \sum_{r,s=1}^{k} \tilde{H}_r^* \tilde{G}_r^* \tilde{F}_{rs} \tilde{G}_s \tilde{H}_s \in \mathbb{R}^{N\times N}$$

is the $(0,0)$-block, which is equal to $\widehat{\Sigma}$. Consider the two matrices $\tilde{W}$ and $\check{W} = \tilde{W} + P_1 + \ldots + P_{2k}$. Then $\mathrm{spec}(\tilde{W}) = \mathrm{spec}(\widehat{\Sigma}) \cup \{0\}$ and $\mathrm{spec}(\check{W}) = \mathrm{spec}(\widehat{\Sigma}) \cup \{1\}$, so

$$\mathrm{spec}(\widehat{\Sigma}) = \mathrm{spec}(\tilde{W}) \cap \mathrm{spec}(\check{W}).$$

Let $X \in \mathbb{R}^{N\times N}$ be a GOE matrix, as in Section 3. Then $\tilde{G}_r$ can be realized as $\tilde{G}_r = \sqrt{\frac{N}{n_r}} P_{r+k} X P_r$. Hence,

$$\tilde{W} = \sum_{r,s=1}^{k} \frac{N}{\sqrt{n_r n_s}} \tilde{H}_r^* P_r X P_{r+k} \tilde{F}_{rs} P_{s+k} X P_s \tilde{H}_s.$$

We construct a free deterministic equivalent in the following way: Let $\mathcal{D} = \langle P_0, \dots, P_{2k} \rangle$, and let $(\mathcal{A}_1, \tau_1)$ be the von Neumann free product of $(\mathcal{D}, N^{-1}\,\mathrm{Tr})$ and a von Neumann probability space containing a semicircular element x. Set $(\mathcal{A}_2, \tau_2) \equiv (\mathbb{C}^{N\times N}, N^{-1}\,\mathrm{Tr})$, which contains $\{\tilde{F}_{rs}, \tilde{H}_r : r, s = 1, \dots, k\}$ and also $\mathcal{D}$. Let $(\mathcal{A}, \tau)$ be the von Neumann amalgamated free product of $(\mathcal{A}_1, \tau_1)$ and $(\mathcal{A}_2, \tau_2)$ with amalgamation over $\mathcal{D}$. In $\mathcal{A}$, identify $f_{rs} \equiv \tilde{F}_{rs}$, $h_r \equiv \tilde{H}_r$, $p_r \equiv P_r$, and define $g_r = \sqrt{N/n_r} p_{r+k} x p_r$. By this construction, x is free of $\mathcal{D}$ (over $\mathbb{C}$) and also free of $\mathcal{A}_2$ over $\mathcal{D}$. Then [NSS02, Proposition 3.7] implies that x is free of $\mathcal{A}_2$ (over $\mathbb{C}$). We may then apply Theorem 3.1 and Assumption 2.2 to conclude

$$\mathrm{spec}(\widehat{\Sigma}) \subset \mathrm{spec}(\tilde{w})_\delta \cap \mathrm{spec}(\check{w})_\delta \tag{5.1}$$

for all large N, where

$$\tilde{w} = \sum_{r,s=1}^{k} \frac{N}{\sqrt{n_r n_s}} h_r^* p_r x p_{r+k} f_{rs} p_{s+k} x p_s h_s = \sum_{r,s=1}^{k} h_r^* g_r^* f_{rs} g_s h_s, \qquad \check{w} = \tilde{w} + p_1 + \ldots + p_{2k}.$$

To finish this proof, we verify that these elements $\{f_{rs}, g_r, h_r, p_r\}$ have the same joint law as described by conditions (1–4) in Section 2.2. Conditions (1–2) are evident by construction. For condition (3), denoting by $\mathrm{NC}_2(2l)$ the non-crossing pairings of $(1, \ldots, 2l)$ and $K(\pi)$ the Kreweras complement of π,

$$\begin{aligned}
\frac{N}{p}\tau\Big((g_r^* g_r)^l\Big) &= \frac{N}{p}\left(\frac{N}{n_r}\right)^l \tau\big((p_r x p_{r+k} x p_r)^l\big) = \frac{N}{p}\left(\frac{N}{n_r}\right)^l \tau\big(x p_{r+k} x p_r \cdots x p_{r+k} x p_r\big) \\
&= \frac{N}{p}\left(\frac{N}{n_r}\right)^l \sum_{\pi \in \mathrm{NC}_2(2l)} \tau_{K(\pi)}\left[p_{r+k}, p_r, \ldots, p_{r+k}, p_r\right] \\
&= \frac{N}{p}\left(\frac{N}{n_r}\right)^l \sum_{m=1}^{l} \tau(p_r)^m \tau(p_{r+k})^{l+1-m} \cdot \left|\{\pi \in \mathrm{NC}_2(2l) : K(\pi) \text{ has } m \text{ blocks of } p_r\}\right| \\
&= \sum_{m=1}^{l} \left(\frac{p}{n_r}\right)^{m-1} \frac{1}{l}\binom{l}{m}\binom{l}{m-1} = \sum_{m=1}^{l} \frac{1}{m}\left(\frac{p}{n_r}\right)^{m-1} \binom{l}{m-1}\binom{l-1}{m-1} = \int x^l \nu_{\frac{p}{n_r}}(x)dx.
\end{aligned}$$

Here, the second line applies [NS06, Theorem 14.4], freeness of $\{p_r, p_{r+k}\}$ and x, and vanishing of all but the second non-crossing cumulant of x. The third line applies $p_r^l = p_r$ and $p_{r+k}^l = p_{r+k}$ for $l \geq 1$, and also that $|K(\pi)| + |\pi| = 2l + 1$ so that $|K(\pi)| = l + 1$. The fourth line applies

$$\left|\{\pi \in \mathrm{NC}_2(2l) : K(\pi) \text{ has } m \text{ blocks of } p_r\}\right| = \left|\{\gamma \in \mathrm{NC}(l) : \gamma \text{ has } m \text{ blocks}\}\right|,$$

which are defined by the Narayana numbers. For more details, see [NS06, Lectures 9, 11, 14]. The last equality is the formula for the l^{th} moment of the Marcenko-Pastur distribution (see [MS17, Exercise 2.11]).

For condition (4), first consider $a_1, \ldots, a_m \in \mathcal{A}_2$ where $a_1, \ldots, a_m$ alternate between the algebras $\langle\{f_{rs}\}, \mathcal{D}\rangle$ and $\langle\{h_r\}, \mathcal{D}\rangle$, and we have $\tau^{\mathcal{D}}(a_i) = 0$ for each i. The latter condition implies that each a_i belonging to $\langle\{f_{rs}\}, \mathcal{D}\rangle$ in fact satisfies $(p_{k+1} + \ldots + p_{2k}) a_i (p_{k+1} + \ldots + p_{2k}) = a_i$, and each a_i belonging to $\langle\{h_r\}, \mathcal{D}\rangle$ in fact satisfies $(p_0 + \ldots + p_k) a_i (p_0 + \ldots + p_k) = a_i$. Then we get $\tau^{\mathcal{D}}(a_1 \ldots a_m) = 0$. This establishes that $\{f_{rs}\}$ and $\{h_r\}$ are free over $\mathcal{D}$. A similar argument shows that $g_1, \ldots, g_k$ are free over $\mathcal{D}$, since each $a_i \in \langle g_r, \mathcal{D}\rangle$ with $\tau^{\mathcal{D}}(a_i) = 0$ must satisfy $(p_r + p_{k+r}) a_i (p_r + p_{k+r}) = a_i$. By construction of the space $\mathcal{A}$, we have that $\{g_1, \ldots, g_k\} \in \mathcal{A}_1$ and $\{f_{rs}, h_r : r, s = 1, \ldots, k\} \in \mathcal{A}_2$ are free over $\mathcal{D}$. Thus condition (4) holds.

Having verified these conditions (1–4), we obtain that μ_0 is the τ^c-law of $\tilde{w}$ in the compressed algebra $\mathcal{A}^c = \{a \in \mathcal{A} : p_0 a p_0 = a\}$ with trace $\tau^c(a) = \tau(p_0)^{-1}\tau(p_0 a p_0)$. Since τ^c is faithful, $\mathrm{supp}(\mu_0)$ is the spectrum of $\tilde{w}$ as an operator in $\mathcal{A}^c$. Then $\mathrm{spec}(\tilde{w}) = \mathrm{supp}(\mu_0) \cup \{0\}$ and $\mathrm{spec}(\check{w}) = \mathrm{supp}(\mu_0) \cup \{1\}$, where $\mathrm{spec}(\cdot)$ here denotes the spectra as operators in $\mathcal{A}$. So $\mathrm{supp}(\mu_0)_\delta = \mathrm{spec}(\tilde{w})_\delta \cap \mathrm{spec}(\check{w})_\delta$ for any $\delta < 1/2$. Combining this with (5.1) concludes the proof. □

Next, we establish the analytic extension of the functions a_r, b_r.

Proof of Proposition 2.5. Denote by $a_r(z)$ and $b_r(z)$ the values of a_r, b_r at $z \in \mathbb{C}^+$, and set $R_0(z) = (z\,\mathrm{Id} + b(z) \cdot \mathring{\Sigma})^{-1}$. Note that $\mathrm{Tr}\, R_0(z) A R_0(z)^* B$ is real and nonnegative for any positive semidefinite A, B. Then from (2.5), we have

$$\begin{aligned}
\mathrm{Im}\, a_r(z) &= -p^{-1}\,\mathrm{Im}\,\mathrm{Tr}\, R_0(z)\mathring{\Sigma}_r \\
&= -p^{-1}\,\mathrm{Im}\,\mathrm{Tr}\left(R_0(z)\mathring{\Sigma}_r R_0(z)^*(z\,\mathrm{Id} + b(z)\cdot\mathring{\Sigma})^*\right) \\
&= p^{-1}(\mathrm{Im}\, z)\,\mathrm{Tr}\, R_0(z)\mathring{\Sigma}_r R_0(z)^* + p^{-1}\sum_{s=1}^{k}(\mathrm{Im}\, b_s(z))\,\mathrm{Tr}\, R_0(z)\mathring{\Sigma}_r R_0(z)^*\mathring{\Sigma}_s.
\end{aligned}$$

In particular, as $\mathrm{Im}\, z > 0$, $\mathrm{Im}\, b_r(z) \geq 0$, and $R_0(z)$ is invertible, we have that either $\mathring{\Sigma}_r = 0$ and $a_r(z) \equiv 0$ for all $z \in \mathbb{C}^+$, or $\mathring{\Sigma}_r \neq 0$ and $\mathrm{Im}\, a_r(z) > 0$ for all $z \in \mathbb{C}^+$. In the former case, a_r trivially extends to $a_r(z) \equiv 0$ on $\mathbb{C} \setminus \mathrm{supp}(\mu_0)$. In the latter case, we recall from the analysis of [FJ16, Theorem 4.1] that each $b_r(iy)$ remains bounded as $y \to \infty$. Then $\lim_{y\to\infty} iy \cdot a_r(iy) = -\,\mathrm{Tr}\,\mathring{\Sigma}_r/m_r$, so $a_r : \mathbb{C}^+ \to \mathbb{C}^+$ is the Stieltjes

transform of a finite measure ν_r on $\mathbb{R}$ with total mass $\nu_r(\mathbb{R}) = \operatorname{Tr} \mathring{\Sigma}_r / m_r$ [GH03, Lemma 2]. Analogous to the above, we also have

$$\operatorname{Im} m_0(z) = p^{-1}(\operatorname{Im} z) \operatorname{Tr} R_0(z) R_0(z)^* + p^{-1} \sum_{s=1}^{k} (\operatorname{Im} b_s(z)) \operatorname{Tr} R_0(z) R_0(z)^* \mathring{\Sigma}_s,$$

and hence for all $z \in \mathbb{C}^+$

$$\operatorname{Im} a_r(z) \le \|\mathring{\Sigma}_r\| \cdot \operatorname{Im} m_0(z).$$

From the Stieltjes inversion formula, this implies $\operatorname{supp}(\nu_r) \subset \operatorname{supp}(\mu_0)$, and hence a_r extends analytically to $\mathbb{C} \setminus \operatorname{supp}(\mu_0)$ also in this case as well.

Then we may extend $b_1(z), \dots, b_k(z)$ to meromorphic functions on $\mathbb{C} \setminus \operatorname{supp}(\mu_0)$ via (2.6), potentially with poles at points $z \in \mathbb{C} \setminus \operatorname{supp}(\mu_0)$ where $\operatorname{Id} + F \operatorname{diag}_n(a(z))$ is singular. We claim that no such points exist: Suppose otherwise, and let $\operatorname{Id} + F \operatorname{diag}_n(a(z_0))$ be singular. Suppose, for notational convenience, that $b_1(z), \dots, b_j(z)$ have poles at z_0, and $b_{j+1}(z), \dots, b_k(z)$ do not. (We may take $j = 0$ or $j = k$ if none or all of the b_r's have poles.) For $z \in \mathbb{C}^-$, it is verified by conjugate-symmetry that $\operatorname{Id} + F \operatorname{diag}_n(a(z))$ is invertible and $\overline{b_r(z)} = b_r(\overline{z})$. Thus $z_0 \in \mathbb{R} \setminus \operatorname{supp}(\mu_0)$. Taking the limit $z \nearrow z_0$ along the real line, and writing as shorthand $D = \operatorname{diag}_n(a(z))$, we have

$$\partial_z \left(-(\operatorname{Id} + FD)^{-1} F \right) = (\operatorname{Id} + FD)^{-1} F \operatorname{diag}(a_1'(z) \operatorname{Id}_{m_1}, \dots, a_k'(z) \operatorname{Id}_{m_k}) (\operatorname{Id} + FD)^{-1} F.$$

Assuming momentarily that F is invertible, $(\operatorname{Id} + FD)^{-1} F = (F^{-1} + D)^{-1}$ is real and symmetric. Then this is also true for non-invertible F by continuity. As each a_s is either identically 0 or the Stieltjes transform of a measure ν_s, we have $a_s'(z) \ge 0$ for all s. Then the above derivative in z is positive-semidefinite. In particular, as $z \nearrow z_0$, each $b_r(z)$ is increasing. So $b_1(z), \dots, b_j(z) \to \infty$ as $z \nearrow z_0$, while $b_{j+1}(z), \dots, b_k(z)$ approach finite values. This implies that for any v in the combined column span of $\mathring{\Sigma}_1, \dots, \mathring{\Sigma}_j$, we have $(z \operatorname{Id} + b \cdot \mathring{\Sigma})^{-1} v \to 0$ as $z \nearrow z_0$. Then $(z \operatorname{Id} + b \cdot \mathring{\Sigma})^{-1} \mathring{\Sigma}_r \to 0$ and $a_r(z_0) = 0$ for each $r = 1, \dots, j$. Denoting by $M(z)$ the lower-right blocks of $\operatorname{Id} + F \operatorname{diag}_n(a(z))$ corresponding to $j+1, \dots, k$, the matrix $\operatorname{Id} + F \operatorname{diag}_n(a(z_0))$ then has the block form

$$\begin{pmatrix} \operatorname{Id} & * \\ 0 & M(z_0) \end{pmatrix}.$$

Since $\operatorname{Id} + F \operatorname{diag}_n(a(z_0))$ is singular, we must have that $M(z_0)$ is singular. Denoting by F_2 the lower-right blocks of F, this implies that $M(z_0) F_2$ is also singular. But the above argument shows $\partial_z(-M(z)F_2)$ is positive-semidefinite, so this must mean $-\operatorname{Tr} M(z) F_2 \to \infty$ as $z \nearrow z_0$. Then $b_r(z) = -m_r^{-1} \operatorname{Tr}_r[M(z)F_2] \to \infty$ for some $r \in \{j+1, \dots, k\}$, contradicting that $b_r(z_0)$ exists and is finite. Thus, $\operatorname{Id} + F \operatorname{diag}_n(a(z))$ is invertible and $b_1, \dots, b_k$ are analytic on all of $\mathbb{C} \setminus \operatorname{supp}(\mu_0)$.

We may then extend $m_0(z)$ to $\mathbb{C} \setminus \operatorname{supp}(\mu_0)$ by (2.7). Note that this must coincide with the Stieltjes transform of μ_0 on $\mathbb{C} \setminus \operatorname{supp}(\mu_0)$, by uniqueness of the analytic extension. Finally, note that if we define $\tilde{a}_1(z), \dots, \tilde{a}_k(z)$ on $\mathbb{C} \setminus \operatorname{supp}(\mu_0)$ by (2.5) from $b_1(z), \dots, b_k(z)$, then each $\tilde{a}_r(z)$ is a meromorphic function on $\mathbb{C} \setminus \operatorname{supp}(\mu_0)$, possibly with poles where $z \operatorname{Id} + b(z) \cdot \mathring{\Sigma}$ is singular. These must agree with $a_1(z), \dots, a_k(z)$ everywhere outside of these poles, as they agree on $\mathbb{C}^+$. Since $a_1(z), \dots, a_k(z)$ are analytic on $\mathbb{C} \setminus \operatorname{supp}(\mu_0)$, no such poles exist, $z \operatorname{Id} + b(z) \cdot \mathring{\Sigma}$ is invertible, and $a_1(z), \dots, a_k(z)$ satisfy (2.5) on all of $\mathbb{C} \setminus \operatorname{supp}(\mu_0)$. □

We record here the following property shown in the above proof.

Proposition 5.1. *For $z \in \mathbb{R} \setminus \operatorname{supp}(\mu_0)$, we have $b_r'(z) \ge 0$ for every $r \in \{1, \dots, k\}$.*

5.2. **Master equation for outlier eigenvalues.** In the remainder of this section, we prove Theorems 2.6 and 2.7. We assume implicitly Assumptions 2.1 and 2.2 throughout. We denote by $C, c > 0$ constants which may change from instance to instance. We fix a constant $\delta > 0$, and define

$$U_\delta = \{z \in \mathbb{C} : \operatorname{dist}(z, \operatorname{supp}(\mu_0)) > \delta\}.$$

For n-dependent matrices $X_1(z), X_2(z)$ of the same dimension, we write

$$X_1(z) \sim X_2(z)$$

if almost surely as $n \to \infty$, we have

$$\sup_{z \in U_\delta} \|X_1(z) - X_2(z)\|_\infty \to 0,$$

where $\|X\|_\infty = \max_{i,j} |X_{i,j}|$.

Following [BGN11], our first step is to establish a "master equation" characterizing outlier eigenvalues of $\widehat{\Sigma}$. Recalling G_r, H_r from (2.8), Ξ_r, Γ_r from (2.1), and F_{rs} from (2.4), we may represent

$$\widehat{\Sigma} = \sum_{r,s=1}^{k} (\Xi_r \Gamma_r + G_r H_r)^\mathsf{T} F_{rs} (\Xi_s \Gamma_s + G_s H_s) = W + P$$

for

$$W = \sum_{r,s=1}^{k} H_r^\mathsf{T} G_r^\mathsf{T} F_{rs} G_s H_s$$
$$P = \sum_{r,s=1}^{k} \Big(\Gamma_r^\mathsf{T} \Xi_r^\mathsf{T} F_{rs} G_s H_s + H_r^\mathsf{T} G_r^\mathsf{T} F_{rs} \Xi_s \Gamma_s + \Gamma_r^\mathsf{T} \Xi_r^\mathsf{T} F_{rs} \Xi_s \Gamma_s \Big).$$

Letting ℓ be the rank of Γ (so $\ell \le \ell_+$), let us write

$$\Gamma = \widetilde{\Gamma} Q^\mathsf{T}$$

where $Q \in \mathbb{R}^{p \times \ell}$ contains the right singular vectors of Γ. We have $Q^\mathsf{T} Q = \mathrm{Id}_\ell$ and $\|\widetilde{\Gamma}\| \le C$. Denote the resolvent of W by

$$R(z) = (W - z\,\mathrm{Id})^{-1}.$$

Define the matrices $\Xi \in \mathbb{R}^{n_+ \times \ell_+}$ and $G \in \mathbb{R}^{n_+ \times kp}$ by the (rectangular) block matrices with non-zero blocks given by

$$\Xi = \begin{bmatrix} \Xi_1 & & & \\ & \Xi_2 & & \\ & & \ddots & \\ & & & \Xi_k \end{bmatrix} \qquad G = \begin{bmatrix} G_1 & & & \\ & G_2 & & \\ & & \ddots & \\ & & & G_k \end{bmatrix}.$$

Finally, define the matrix $H \in \mathbb{R}^{kp \times p}$ as the vertical stacking of $\{H_r\}_{r=1}^k$, and set

$$S(z) = \Xi^\mathsf{T} F G H R(z) Q.$$

Define the matrix

$$\widehat{K}(z) = \mathrm{Id} + \begin{bmatrix} S(z) \cdot \widetilde{\Gamma}^\mathsf{T} & \Xi^\mathsf{T} F G H R(z) H^\mathsf{T} G^\mathsf{T} F \Xi \cdot \widetilde{\Gamma} + S(z) \cdot \widetilde{\Gamma}^\mathsf{T} \Xi^\mathsf{T} F \Xi \widetilde{\Gamma} \\ Q^\mathsf{T} R(z) Q \cdot \widetilde{\Gamma}^\mathsf{T} & S(z)^\mathsf{T} \cdot \widetilde{\Gamma} + Q^\mathsf{T} R(z) Q \cdot \widetilde{\Gamma}^\mathsf{T} \Xi^\mathsf{T} F \Xi \widetilde{\Gamma} \end{bmatrix}. \tag{5.2}$$

Lemma 5.2. *The eigenvalues of $\widehat{\Sigma}$ which are not eigenvalues of W are the roots of* $\det \widehat{K}(z) = 0$.

Proof. The eigenvalues of $\widehat{\Sigma}$ which are not eigenvalues of W are the roots of

$$\det\Big(R(z)(\widehat{\Sigma} - z\,\mathrm{Id}) \Big) = 0.$$

In the above notation, we have

$$P = Q \widetilde{\Gamma}^\mathsf{T} \Xi^\mathsf{T} F G H + H^\mathsf{T} G^\mathsf{T} F \Xi \widetilde{\Gamma} Q^\mathsf{T} + Q \widetilde{\Gamma}^\mathsf{T} \Xi^\mathsf{T} F \Xi \widetilde{\Gamma} Q^\mathsf{T},$$

from which we may compute

$$\begin{aligned} R(z)(\widehat{\Sigma} - z\,\mathrm{Id}) &= \mathrm{Id} + R(z) P \\ &= \mathrm{Id} + \begin{bmatrix} R(z) Q \widetilde{\Gamma}^\mathsf{T} & R(z) H^\mathsf{T} G^\mathsf{T} F \Xi \widetilde{\Gamma} + R(z) Q \widetilde{\Gamma}^\mathsf{T} \Xi^\mathsf{T} F \Xi \widetilde{\Gamma} \end{bmatrix} \begin{bmatrix} \Xi^\mathsf{T} F G H \\ Q^\mathsf{T} \end{bmatrix}. \end{aligned}$$

Applying the identity $\det(\mathrm{Id} + XY) = \det(\mathrm{Id} + YX)$, we find that

$$0 = \det\left(\mathrm{Id} + \begin{bmatrix} \Xi^\mathsf{T} F G H \\ Q^\mathsf{T} \end{bmatrix} \begin{bmatrix} R(z) Q \widetilde{\Gamma}^\mathsf{T} & R(z) H^\mathsf{T} G^\mathsf{T} F \Xi \widetilde{\Gamma} + R(z) Q \widetilde{\Gamma}^\mathsf{T} \Xi^\mathsf{T} F \Xi \widetilde{\Gamma} \end{bmatrix} \right) = \det \widehat{K}(z). \qquad \square$$

5.3. **Approximation by trace.** Next, we apply concentration of measure over the randomness of Ξ to show that each (block) matrix element $\{\widehat{K}_{ij}(z)\}_{i,j=1}^2$ of $\widehat{K}(z)$ is well-approximated by certain traces. Define the Schur complement

$$\widehat{T}(z) = \widehat{K}_{22}(z) - \widehat{K}_{21}(z)\widehat{K}_{11}(z)^{-1}\widehat{K}_{12}(z) \tag{5.3}$$

and the matrix

$$\check{T}(z) = \mathrm{Id} + Q^{\mathsf{T}}R(z)Q \cdot \widetilde{\Gamma}^{\mathsf{T}} \left(\sum_{r=1}^{k} n_r^{-1} \operatorname{Tr}_r[F - FGHR(z)H^{\mathsf{T}}G^{\mathsf{T}}F] \, \mathrm{Id}_{\ell_r} \right) \widetilde{\Gamma}. \tag{5.4}$$

Lemma 5.3. *We have that $S(z) \sim 0$, $\widehat{K}_{11}(z) \sim \mathrm{Id}_{\ell_+}$, and $\widehat{T}(z) \sim \check{T}(z)$.*

Our proof of Lemma 5.3 will apply the following concentration result, from [BEK+14, Lemma 3.1].

Proposition 5.4 ([BEK+14, Lemma 3.1])**.** *Let $x, y \in \mathbb{R}^N$ be independent vectors with independent entries satisfying*

$$\mathbb{E}[x_i] = \mathbb{E}[y_i] = 0, \qquad \mathbb{E}[x_i^2] = \mathbb{E}[y_i^2] = 1/N, \qquad \mathbb{E}[|x_i|^k] < C_k N^{-k/2}, \qquad \mathbb{E}[|y_i|^k] < C_k N^{-k/2}$$

for each $k \geq 1$ and some constants $C_k > 0$. Let $A_1, A_2 \in \mathbb{C}^{N \times N}$ be any deterministic matrices and $v \in \mathbb{C}^N$ any deterministic vector. Then for any $\tau, D > 0$ and all $N \geq N_0(\tau, D)$,

$$\mathbb{P}[|x^{\mathsf{T}}v| \geq N^{-1/2+\tau}\|v\|_2] < N^{-D},$$

$$\mathbb{P}[|x^{\mathsf{T}}A_1x - \operatorname{Tr} A_1| \geq N^{-1+\tau}\|A_1\|_{HS}] < N^{-D}, \quad \mathbb{P}[|x^{\mathsf{T}}A_2y| \geq N^{-1+\tau}\|A_2\|_{HS}] < N^{-D}.$$

For a sufficiently large constant $C > 0$, define the good event

$$\mathcal{E}_n = \{\operatorname{spec}(W) \subset \operatorname{supp}(\mu_0)_{\delta/2}, \ \|G_r\| < C, \ \|\Xi_r\| < C \text{ for all } r = 1, \dots, k\}. \tag{5.5}$$

From Theorem 2.4 and Assumption 2.1, we have that $\mathcal{E}_n$ holds almost surely for all large n. We will use implicitly throughout that on this event $\mathcal{E}_n$, we have $\|G\| < C$, $\|\Xi\| < C$, $\|R(z)\| < C\min(1, 1/|z|)$, and $\|R'(z)\| < C\min(1, 1/|z|^2)$ for all $z \in U_\delta$ and a constant $C > 0$.

Proof of Lemma 5.3. Note that $S(z)$ has blocks given by

$$\sum_{s=1}^{k} \Xi_r^{\mathsf{T}} F_{rs} G_s H_s R(z) Q$$

for $r = 1, \dots, k$, where $\Xi_1, \dots, \Xi_k$ are independent of $G_1, \dots, G_k$. On the event $\mathcal{E}_n$, for any fixed $\varepsilon > 0$, we have $\|S(z)\|_\infty < \varepsilon$ for all $|z| > C_0$ and some constant $C_0 > 0$. For $|z| \leq C_0$, note that $\|F_{rs}G_sH_sR(z)Q\| < C$ for all $z \in U_\delta$. Then this bound holds for the ℓ_2-norm of each column of $F_{rs}G_sH_sR(z)Q$. The entries of Ξ_r satisfy the conditions of Proposition 5.4 with $N = n_r$. Applying the proposition conditional on $G_1, \dots, G_k$ and on $\mathcal{E}_n$, we get $\|\Xi_r^{\mathsf{T}}F_{rs}G_sH_sR(z)Q\|_\infty < n^{-1/2+\tau}$ with probability $1 - n^{-D}$, and hence $\|S(z)\|_\infty < n^{-1/2+\tau}$ as well. Taking a union bound over a grid of values in $U_\delta \cap \{|z| \leq C_0\}$ with spacing $n^{-1/2}$, and applying Lipschitz continuity of $S(z)$ on $\mathcal{E}_n$, we get almost surely

$$\sup_{z \in U_\delta : |z| \leq C_0} \|S(z)\|_\infty \to 0.$$

Then $\limsup_{n\to\infty} \sup_{z\in U_\delta} \|S(z)\|_\infty \leq \varepsilon$. As $\varepsilon > 0$ is arbitrary, this shows $S(z) \sim 0$. This implies also $\widehat{K}_{11}(z) \sim \mathrm{Id}_{\ell_+}$.

For $\widehat{T}(z)$, note first that $S(z) \sim 0$ and $\widehat{K}_{11}(z) \sim \mathrm{Id}_{\ell_+}$ imply

$$\widehat{T}(z) \sim \mathrm{Id} + Q^{\mathsf{T}}R(z)Q \cdot \widetilde{\Gamma}^{\mathsf{T}}\Xi^{\mathsf{T}}\Big(F - FGHR(z)H^{\mathsf{T}}G^{\mathsf{T}}F\Big)\Xi\widetilde{\Gamma}.$$

Notice that $\Xi^{\mathsf{T}}\Big(F - FGHR(z)H^{\mathsf{T}}G^{\mathsf{T}}F\Big)\Xi$ is a $k \times k$ block matrix with blocks

$$\Xi_r^{\mathsf{T}} Y_{rs}(z) \Xi_s, \qquad Y_{rs}(z) = F_{rs} - \sum_{r',s'=1}^{k} F_{rs'}G_{s'}H_{s'}R(z)H_{r'}^{\mathsf{T}}G_{r'}^{\mathsf{T}}F_{r's}.$$

On $\mathcal{E}_n$, we bound $\|Y_{rs}(z)\|_{\mathrm{HS}} \le C\sqrt{n}\|Y_{rs}(z)\| \le C'\sqrt{n}$. Then, applying Proposition 5.4 again for each pair (r,s) and each pair of columns of Ξ_r and Ξ_s, we get for each fixed $z \in U_\delta$ that

$$\big\|\Xi_r^{\mathsf{T}} Y_{rs}(z)\Xi_s - \mathbf{1}\{r=s\}n_r^{-1}\operatorname{Tr}_r[F - FGHR(z)H^{\mathsf{T}}G^{\mathsf{T}}F]\cdot \mathrm{Id}_{\ell_r}\big\|_\infty < n^{-1/2+\tau}$$

with probability $1-n^{-D}$. Applying Lipschitz continuity and a union bound over a grid of values $|z| \le C_0$, a separate argument for $|z| > C_0$ as above, and the Borel-Cantelli lemma, we obtain the lemma. □

5.4. **Approximation by deterministic equivalents.** We now approximate the terms appearing in (5.4) by quantities in the free deterministic equivalent model described in Section 2.2. Define

$$\widetilde{T}(z) = \mathrm{Id} + Q^{\mathsf{T}}(z\,\mathrm{Id} + b\cdot\mathring{\Sigma})^{-1}Q\cdot\widetilde{\Gamma}^{\mathsf{T}}\operatorname{diag}_\ell(b)\widetilde{\Gamma},$$

where $\widetilde{\Gamma}^{\mathsf{T}}\operatorname{diag}_\ell(b)\widetilde{\Gamma} = \sum_{r=1}^k b_r\widetilde{\Gamma}_r^{\mathsf{T}}\widetilde{\Gamma}_r$. By Proposition 2.5, $\widetilde{T}(z)$ is a well-defined analytic function on $\mathbb{C}\setminus\operatorname{supp}(\mu_0)$. We establish the following approximation, which follows immediately from the definitions of $\check{T}$, $\widetilde{T}$, and the estimates of the next two propositions.

Lemma 5.5. *We have $\check{T}(z) \sim \widetilde{T}(z)$.*

Proposition 5.6. *We have $Q^{\mathsf{T}}R(z)Q \sim -Q^{\mathsf{T}}(z\,\mathrm{Id} + b\cdot\mathring{\Sigma})^{-1}Q$.*

Proof. The von Neumann probability space $(\mathcal{A},\tau)$ in Section 2.2 may be constructed in the following way: Let $(\mathcal{A}_1,\tau_1) = (\mathbb{C}^{N\times N}, N^{-1}\operatorname{Tr})$, containing the embedded matrices $\tilde{H}_1,\dots,\tilde{H}_k$ and $P_0,\dots,P_{2k}$. Identify $h_r = \tilde{H}_r$ and $p_r = P_r$. Construct a von Neumann probability space $(\mathcal{A}_2,\tau_2)$ also containing $p_0,\dots,p_{2k}$ and elements $\{f_{rs}, g_r : r,s = 1,\dots,k\}$ satisfying all required conditions on their joint law under τ_2. Let $(\mathcal{A},\tau)$ be the von Neumann amalgamated free product over $\langle p_0,\dots,p_{2k}\rangle$.

Let $w = \sum_{r,s} h_r^* g_r^* f_{rs} g_s h_s \in \mathcal{A}$. By Corollary 4.3 applied to each pair of columns of Q, we find that

$$Q^{\mathsf{T}}R(z)Q \sim Q^{\mathsf{T}}P_0\tau^{\mathcal{H}}((w-z)^{-1})P_0Q$$

where $P_0\tau^{\mathcal{H}}((w-z)^{-1})P_0$ is identified with its upper-left block as an element of $\mathbb{C}^{p\times p}$. By [FJ16, Equation (4.12)] and the identification $\beta_r(z) = -b_r(z)$ at the conclusion of the proof of [FJ16, Lemma 4.4], we have

$$\tau^{\mathcal{H}}\Big((w-z)^{-1}\Big) = -\left(z + \sum_{r=1}^k h_r^* h_r b_r(z)\right)^{-1}.$$

Recalling the identification $h_r = \tilde{H}_r$ and interpreting this as an element of $\mathbb{C}^{N\times N}$, its upper-left block is

$$-\left(z + \sum_{r=1}^k H_r^{\mathsf{T}} H_r b_r(z)\right)^{-1} = -(z + b\cdot\mathring{\Sigma})^{-1},$$

which concludes the proof. □

Proposition 5.7. *Fix $\delta > 0$. Almost surely as $n\to\infty$, for each $t \in \{1,\dots,k\}$, we have*

$$\sup_{z\in U_\delta}\big|n_t^{-1}\operatorname{Tr}_t[F - FGHR(z)H^{\mathsf{T}}G^{\mathsf{T}}F] + b_t(z)\big| \to 0.$$

The rest of this section is devoted to the proof of Proposition 5.7. In the von Neumann probability space $(\mathcal{A},\tau)$ defined in Section 2.2, let $\mathcal{H} = \langle h_1,\dots,h_k\rangle$, $\mathcal{G} = \langle g_1,\dots,g_k\rangle$, $\mathcal{F} = \langle f_{11}, f_{12},\dots,f_{kk}\rangle$ and $\mathcal{D} = \langle p_0,\dots,p_{2k}\rangle$ be the generated von Neumann subalgebras of $\mathcal{A}$. Define the elements

$$w = \sum_{r,s=1}^k h_r^* g_r^* f_{rs} g_s h_s \qquad v = \sum_{r,s=1}^k g_r^* f_{rs} g_s \qquad u = \sum_{r,s=1}^k f_{rs}. \tag{5.6}$$

For any $r,s,t \in \{1,\dots,k\}$ define

$$a_{rts} = h_r^* g_r^* f_{rt} f_{ts} g_s h_s \qquad b_{rts} = g_r^* f_{rt} f_{ts} g_s \qquad c_{rts} = f_{rt} f_{ts}.$$

Our goal is to compute

$$\sum_{r,s=1}^k \tau(f_{ts} g_s h_s (w-z)^{-1} h_r^* g_r^* f_{rt}) = \sum_{s,t=1}^k \tau(a_{rts}(w-z)^{-1}),$$

and we will do this using the left-augmented Cauchy- and $\mathcal{R}$-transforms introduced in Section 3.1.

Recall the $\mathcal{H}$-valued conditional expectation $\tau^{\mathcal{H}}$, Cauchy-transform $G^{\mathcal{H}}$, and $\mathcal{R}$-transform $\mathcal{R}^{\mathcal{H}}$ from Section 3.1, and similarly for $\mathcal{G}$ and $\mathcal{D}$. For each $i \in \{0, \dots, 2k\}$, denote

$$\tau_i(a) = \tau(p_i)^{-1}\tau(p_i a p_i)$$

and note that

$$\tau^{\mathcal{D}}(a) = \sum_{i=0}^{2k} \tau_i(a) p_i.$$

For a sufficiently large constant $C > 0$, define

$$\mathbb{D} = \{z \in \mathbb{C} : |z| > C\}.$$

We define the following analytic functions $\{\alpha_i\}_{i=0}^{2k}$, $\{\beta_i\}_{i=0}^{2k}$, $\{d_i\}_{i=0}^{2k}$, $\{\gamma_j\}_{j=0}^{2k}$, $\{\delta_j\}_{j=0}^{2k}$, and $\{e_j\}_{j=0}^{2k}$ on $\mathbb{D}$, also used in [FJ16]: Define $\{\alpha_i\}_{i=0}^{2k}$ and $\{\beta_i\}_{i=0}^{2k}$ by

$$\alpha_i = \tau_i(h_i G_w^{\mathcal{H}}(z) h_i^*) \qquad \beta_i = \tau_i\left(R_v^{\mathcal{D}}\left(\sum_{i=1}^{k} \alpha_i p_i\right)\right) \qquad \text{for } i \in \{1, \dots, k\} \tag{5.7}$$

and $\alpha_0 = \alpha_{k+1} = \cdots = \alpha_{2k} = |z|^{-1}$ and $\beta_0 = \beta_{k+1} = \cdots = \beta_{2k} = 0$. Then set

$$d_i = \alpha_i^{-1} + \beta_i, \quad d = \sum_{i=0}^{2k} d_i p_i.$$

Now define $\{\gamma_j\}_{j=0}^{2k}$ and $\{\delta_j\}_{j=0}^{2k}$ by

$$\gamma_{j+k} = \tau_{j+k}(g_j G_v^{\mathcal{G}}(d) g_j^*) \qquad \delta_{j+k} = \tau_{j+k}\left(R_u^{\mathcal{D}}\left(\sum_{j=k+1}^{2k} \gamma_j p_j\right)\right) \qquad \text{for } j \in \{1, \dots, k\} \tag{5.8}$$

and $\gamma_0 = \gamma_1 = \cdots = \gamma_k = |z|^{-1}$ and $\delta_0 = \delta_1 = \cdots = \delta_k = 0$. Finally, set

$$e_j = \gamma_j^{-1} + \delta_j, \quad e = \sum_{j=0}^{2k} e_j p_j.$$

These quantities satisfy the following identities, all shown in [FJ16].

Proposition 5.8. *The following statements hold for $\alpha_i, \beta_i, d_i, \gamma_j, \delta_j, e_j$ on $\mathbb{D}$.*

(a) $\sum_{i=0}^{2k} \alpha_i p_i = G_v^{\mathcal{D}}(d)$.
(b) $\sum_{j=0}^{2k} \gamma_j p_j = G_u^{\mathcal{D}}(e)$.
(c) *The quantities $a_r = -\frac{p\alpha_r}{n_r}$ and $b_r = -\beta_r$ satisfy the relations (2.5) and (2.6).*
(d) *For $r \in \{1, \dots, k\}$, we have $e_{r+k} = -a_r^{-1}$.*

Proof. (a) follows from [FJ16, Equation (4.15)], (b) follows from [FJ16, Equation (4.21)], (c) is shown at the end of the proof of [FJ16, Lemma 4.4], and (d) follows from [FJ16, Equation (4.28)]. □

Proposition 5.9. *We have*

$$R_{a_{rts},w}^{\mathcal{H}}(G_w^{\mathcal{H}}(z)) = h_r^* h_r \tau_r\left[R_{b_{rts},v}^{\mathcal{D}}\left(G_v^{\mathcal{D}}(d)\right)\right], \qquad R_{b_{rts},v}^{\mathcal{G}}(G_v^{\mathcal{G}}(d)) = g_r^* g_r \tau_{r+k}\left[R_{c_{rts},u}^{\mathcal{D}}\left(G_u^{\mathcal{D}}(e)\right)\right].$$

Proof. For the first equality, notice that for $c = G_w^{\mathcal{H}}(z)$, we have

$$\begin{aligned}
\kappa_l^{\mathcal{H}}(a_{rts}, cw, \dots, cw) &= \sum_{\substack{r_2,\dots,r_l=1 \\ s_2,\dots,s_l=1}}^{k} \kappa_l^{\mathcal{H}}\left(h_r^* g_r^* f_{rt} f_{ts} g_s h_s, ch_{r_2}^* g_{r_2}^* f_{r_2 s_2} g_{s_2} h_{s_2}, \dots, ch_{r_l}^* g_{r_l}^* f_{r_l s_l} g_{s_l} h_{s_l}\right) \\
&= \sum_{\substack{r_2,\dots,r_l=1 \\ s_2,\dots,s_l=1}}^{k} h_r^* \kappa_l^{\mathcal{H}}\left(g_r^* f_{rt} f_{ts} g_s, h_s ch_{r_2}^* g_{r_2}^* f_{r_2 s_2} g_{s_2}, \dots, h_{s_{l-1}} ch_{r_l}^* g_{r_l}^* f_{r_l s_l} g_{s_l}\right) h_{s_l} \\
&= \sum_{\substack{r_2,\dots,r_l=1 \\ s_2,\dots,s_l=1}}^{k} h_r^* \kappa_l^{\mathcal{D}}\left(g_r^* f_{rt} f_{ts} g_s, \tau^{\mathcal{D}}(h_s ch_{r_2}^*) g_{r_2}^* f_{r_2 s_2} g_{s_2}, \dots, \tau^{\mathcal{D}}(h_{s_{l-1}} ch_{r_l}^*) g_{r_l}^* f_{r_l s_l} g_{s_l}\right) h_{s_l},
\end{aligned}$$

where we apply [NSS02, Theorem 3.6] and $\mathcal{D}$-freeness of $\{\mathcal{F},\mathcal{G}\}$ and $\mathcal{H}$ in the last step. Notice now that $\tau^{\mathcal{D}}(h_s c h_r^*)=0$ unless $s=r$, that for any $d'\in\mathcal{D}$ we have $h_r^* d' h_r = h_r^* h_r \tau_r(d')$, and that

$$\tau^{\mathcal{D}}(h_r c h_r^*) g_r^* = \tau_r(h_r c h_r^*) p_r g_r^* = \Big(\sum_{i=0}^{2k} \alpha_i p_i\Big) g_r^*.$$

Therefore, applying Proposition 5.8(a) and defining $c'=G_v^{\mathcal{D}}(d)$, we have that

$$\kappa_l^{\mathcal{H}}(a_{rts}, cw, \dots, cw) = h_r^* h_r \sum_{r_3,\dots,r_l=1}^{k} \tau_r\Big(\kappa_l^{\mathcal{D}}\Big(g_r^* f_{rt} f_{ts} g_s, c' g_s^* f_{sr_3} g_{r_3}, c' g_{r_3}^* f_{r_3 r_4} g_{r_4}, \dots, c' g_{r_l}^* f_{r_l r} g_r\Big)\Big).$$

On the other hand, using $g_s = g_s p_s$ and $p_s c' p_r = 0$ unless $s=r$, we have

$$\begin{aligned}
\kappa_l^{\mathcal{D}}(b_{rts}, c'v, \dots, c'v) &= \sum_{\substack{r_2,\dots,r_l=1\\ s_2,\dots,s_l=1}}^{k} \kappa_l^{\mathcal{D}}\Big(g_r^* f_{rt} f_{ts} g_s, c' g_{r_2}^* f_{r_2 s_2} g_{s_2}, \dots, c' g_{r_l}^* f_{r_l s_l} g_{s_l}\Big) \\
&= \sum_{\substack{r_2,\dots,r_l=1\\ s_2,\dots,s_l=1}}^{k} \kappa_l^{\mathcal{D}}\Big(g_r^* f_{rt} f_{ts} g_s, p_s c' p_{r_2} g_{r_2}^* f_{r_2 s_2} g_{s_2}, \dots, p_{s_{l-1}} c' p_{r_l} g_{r_l}^* f_{r_l s_l} g_{s_l}\Big) \\
&= \sum_{r_3,\dots,r_l=1}^{k} \kappa_l^{\mathcal{D}}\Big(g_r^* f_{rt} f_{ts} g_s, c' g_s^* f_{sr_3} g_{r_3}, c' g_{r_3}^* f_{r_3 r_4} g_{r_4}, \dots, c' g_{r_l}^* f_{r_l r} g_r\Big).
\end{aligned}$$

Comparing with the above,

$$\kappa_l^{\mathcal{H}}(a_{rts}, cw, \dots, cw) = h_r^* h_r \tau_r\Big(\kappa_l^{\mathcal{D}}(b_{rts}, c'v, \dots, c'v)\Big).$$

Summing over $l \geq 1$ yields the first identity. The proof of the second identity is exactly parallel, using Proposition 5.8(b) in place of Proposition 5.8(a). □

Proposition 5.10. *We have*

$$\tau(a_{rts}(z-w)^{-1}) = \tau\Big(f_{rt}(e-u)^{-1} f_{ts}\Big).$$

Proof. Notice first that since $a_{rts}(z-w)^{-1} = p_0 a_{rts}(z-w)^{-1}$,

$$\tau(a_{rts}(z-w)^{-1}) = \tau\left(G^{\mathcal{H}}_{a_{rts},w}(z)\right) = \tau(p_0)\tau_0\left(G^{\mathcal{H}}_{a_{rts},w}(z)\right).$$

Substituting the expression of Proposition 5.9 into the identity

$$G^{\mathcal{H}}_{a_{rts},w}(z) = R^{\mathcal{H}}_{a_{rts},w}(G^{\mathcal{H}}_w(z)) G^{\mathcal{H}}_w(z)$$

of Lemma 3.3, we find that

$$G^{\mathcal{H}}_{a_{rts},w}(z) = h_r^* h_r \cdot G^{\mathcal{H}}_w(z) \tau_r\left[R^{\mathcal{D}}_{b_{rts},v}\Big(G^{\mathcal{D}}_v(d)\Big)\right],$$

from which we obtain

$$\tau_0[G^{\mathcal{H}}_{a_{rts},w}(z)] = \tau_0[h_r^* h_r G^{\mathcal{H}}_w(z)] \tau_r\left[R^{\mathcal{D}}_{b_{rts},v}\Big(G^{\mathcal{D}}_v(d)\Big)\right].$$

Noting that $\tau_0[h_r^* h_r G^{\mathcal{H}}_w(z)] = \frac{\tau(p_r)}{\tau(p_0)}\alpha_r$, we obtain

$$\begin{aligned}
\tau_0[G^{\mathcal{H}}_{a_{rts},w}(z)] &= \frac{\tau(p_r)}{\tau(p_0)} \tau_r\left[R^{\mathcal{D}}_{b_{rts},v}(G^{\mathcal{D}}_v(d))\alpha_r\right] \\
&= \frac{\tau(p_r)}{\tau(p_0)} \tau_r\left[R^{\mathcal{D}}_{b_{rts},v}(G^{\mathcal{D}}_v(d)) G^{\mathcal{D}}_v(d)\right] = \frac{\tau(p_r)}{\tau(p_0)} \tau_r[G^{\mathcal{D}}_{b_{rts},v}(d)] = \frac{\tau(p_r)}{\tau(p_0)} \tau_r[G^{\mathcal{G}}_{b_{rts},v}(d)],
\end{aligned}$$

where in the second equality we replace α_r by $G^{\mathcal{D}}_v(d) = \sum_{i=0}^{2k} \alpha_i p_i$. Substituting Proposition 5.9 into the identity

$$G^{\mathcal{G}}_{b_{rts},v}(d) = R^{\mathcal{G}}_{b_{rts},v}(G^{\mathcal{G}}_v(d)) G^{\mathcal{G}}_v(d),$$

we find that

$$G^{\mathcal{G}}_{b_{rts},v}(d) = g_r^* g_r G^{\mathcal{G}}_v(d) \tau_{r+k}\left[R^{\mathcal{D}}_{c_{rts},u}\Big(G^{\mathcal{D}}_u(e)\Big)\right].$$

Noting that $\tau_r(g_r^* g_r G_v^{\mathcal{G}}(d)) = \frac{\tau(p_{r+k})}{\tau(p_r)}\gamma_{r+k}$, we find similarly that

$$\tau_r[G^{\mathcal{G}}_{b_{rts},v}(d)] = \frac{\tau(p_{r+k})}{\tau(p_r)}\tau_{r+k}\left[R^{\mathcal{D}}_{c_{rts},u}\Big(G_u^{\mathcal{D}}(e)\Big)\gamma_{r+k}\right] = \frac{\tau(p_{r+k})}{\tau(p_r)}\tau_{r+k}\left[R^{\mathcal{D}}_{c_{rts},u}\Big(G_u^{\mathcal{D}}(e)\Big)G_u^{\mathcal{D}}(e)\right] = \frac{\tau(p_{r+k})}{\tau(p_r)}\tau_{r+k}\left[G^{\mathcal{D}}_{c_{rts},u}\Big(e\Big)\right].$$

Putting everything together, we conclude that

$$\tau(a_{rts}(z-w)^{-1}) = \tau(p_{r+k})\tau_{r+k}\left[G^{\mathcal{D}}_{c_{rts},u}\Big(e\Big)\right] = \tau\Big(f_{rt}f_{ts}(e-u)^{-1}\Big) = \tau\Big(f_{ts}(e-u)^{-1}f_{rt}\Big). \qquad \square$$

Proof of Proposition 5.7. For any $\varepsilon > 0$, we may choose $K > 0$ so that almost surely for all large n,

$$\sup_{z\in\mathbb{D}}\left\|\sum_{l=K+1}^{\infty} z^{-l-1}W^l\right\| < \varepsilon, \qquad \sup_{z\in\mathbb{D}}\left\|\sum_{l=K+1}^{\infty} z^{-l-1}w^l\right\| < \varepsilon.$$

Then applying the convergent series expansions of $-R(z) = (z-W)^{-1}$ and $(z-w)^{-1}$ on $\mathbb{D}$, the fact that $\{H_r\}_{r=1}^k$, $\{G_r\}_{r=1}^k$, and $\{F_{rs}\}_{r,s=1}^k$ are almost surely uniformly bounded for large n, and the conclusion

$$\tau(a_{rts}w^l) - N^{-1}\operatorname{Tr} H_r^{\mathsf{T}}G_r^{\mathsf{T}}F_{rt}F_{ts}G_sH_sW^l \to 0$$

for each fixed $l \in \{0,\dots,K\}$ by [FJS18, Theorem 3.9], we obtain

$$\sup_{z\in\mathbb{D}}\left|-N^{-1}\operatorname{Tr}[H_r^{\mathsf{T}}G_r^{\mathsf{T}}F_{rt}F_{ts}G_sH_sR(z)] - \tau(a_{rts}(z-w)^{-1})\right| < 2\varepsilon.$$

As $\varepsilon > 0$ is arbitrary, the left side converges to 0 almost surely. By Lemma 4.7, we may then replace the supremum over $\mathcal{D}$ with one over U_δ. Applying Proposition 5.10, we find that

$$\frac{1}{n_t}\operatorname{Tr}_t[FGHR(z)H^{\mathsf{T}}G^{\mathsf{T}}F] = \sum_{r,s=1}^k \frac{1}{n_t}\operatorname{Tr}[H_r^{\mathsf{T}}G_r^{\mathsf{T}}F_{rt}F_{ts}G_sH_sR(z)]$$
$$\sim \sum_{r,s=1}^k -\frac{N}{n_t}\tau(a_{rts}(z-w)^{-1}) = \sum_{r,s=1}^k -\frac{N}{n_t}\tau\Big(f_{ts}(e-u)^{-1}f_{rt}\Big) = \frac{1}{n_t}\operatorname{Tr}_t\Big(F(\operatorname{diag}_n(a^{-1})+F)^{-1}F\Big),$$

the last line applying equality in law of $\{\tilde{F}_{rs}\}$ with $\{f_{rs}\}$, the definitions of e and u, and Proposition 5.8(d). Notice now that by the Woodbury identity,

$$F - F(\operatorname{diag}_n(a^{-1})+F)^{-1}F = (F^{-1}+\operatorname{diag}_n(a))^{-1} = (\mathrm{Id}+F\operatorname{diag}_n(a))^{-1}F,$$

which holds also for non-invertible F by continuity. Taking the block trace Tr_t and comparing with the definition of b_t in (2.6) concludes the proof. $\square$

5.5. **Outlier eigenvalues.** We now prove Theorem 2.6 on the outlier eigenvalues.

Proposition 5.11. *There is a constant $C > 0$ such that for all $z \in U_\delta$, $r \in \{1,\dots,k\}$, and large enough n, we have $|b_r(z)| < C$.*

Proof. By Proposition 5.7, almost surely as $n \to \infty$ we have

$$\sup_{z\in U_\delta}\left|n_r^{-1}\operatorname{Tr}_r[FGHR(z)H^{\mathsf{T}}G^{\mathsf{T}}F - F] - b_r(z)\right| \to 0.$$

On the event $\mathcal{E}_n$ of (5.5), by Assumption 2.2, we see that for each $z \in U_\delta$,

$$\left\|FGHR(z)H^{\mathsf{T}}G^{\mathsf{T}}F - F\right\| < C,$$

and hence $|b_r(z)| < C$ almost surely for large n. Then this holds deterministically for large n, since $b_r(z)$ is deterministic. $\square$

Proposition 5.12. *There is a constant $C > 0$ such that* $\operatorname{supp}(\mu_0) \subset [-C, C]$.

Proof. The law μ_0 is the τ^c-distribution of $w = \sum_{r,s=1}^k h_r^* g_r^* f_{rs} g_s h_s$ in the compressed algebra $(\mathcal{A}^c, \tau^c)$. We have $\|w\| \le C$ for a constant $C > 0$, hence $\operatorname{supp}(\mu_0) = \operatorname{spec}(w) \subset [-C, C]$. $\square$

Proposition 5.13. *The following properties hold for $\widetilde{T}(z)$ and all large n.*

(a) *All roots of* $\det \widetilde{T}(z) = 0$ *in* $\mathbb{C} \setminus \operatorname{supp}(\mu_0)$ *are real.*
(b) *There exists some* $R > 0$ *so that all roots of* $\det \widetilde{T}(z) = 0$ *have magnitude at most* R*.*
(c) *For* $\delta > 0$*, there is a constant* $C > 0$ *such that*

$$\sup_{z \in U_\delta} \|\widetilde{T}(z)\|_\infty < C, \qquad \sup_{z \in U_\delta} |\det \widetilde{T}(z)| < C.$$

The following properties hold for $\widehat{T}(z)$ *almost surely for all large* n*.*

(a') *For* $\delta > 0$*, all roots of* $\det \widehat{T}(z) = 0$ *in* U_δ *are real.*
(b') *There exists some* $R > 0$ *so that all roots of* $\det \widehat{T}(z) = 0$ *in* U_δ *have magnitude at most* R*.*
(c') *For* $\delta > 0$*, there is a constant* $C > 0$ *such that we have*

$$\sup_{z \in U_\delta} \|\widehat{T}(z)\|_\infty < C, \qquad \sup_{z \in U_\delta} |\det \widehat{T}(z)| < C.$$

Proof. We first prove the statements for $\widetilde{T}(z)$. For (a), applying $\det(\mathrm{Id} + XY) = \det(\mathrm{Id} + YX)$ and the fact that $z \,\mathrm{Id} + b \cdot \mathring{\Sigma}$ is invertible for $z \in \mathbb{C} \setminus \operatorname{supp}(\mu_0)$ from Proposition 2.5, we have

$$\begin{aligned} 0 = \det \widetilde{T}(z) &\Leftrightarrow 0 = \det \Big(\mathrm{Id} + (z \,\mathrm{Id} + b \cdot \mathring{\Sigma})^{-1} \Gamma^\mathsf{T} \operatorname{diag}_\ell(b) \Gamma \Big) \\ &\Leftrightarrow 0 = \det \Big(z \,\mathrm{Id} + b \cdot \mathring{\Sigma} + \Gamma^\mathsf{T} \operatorname{diag}_\ell(b) \Gamma \Big). \end{aligned} \tag{5.9}$$

For $z \in \mathbb{C}^+$ and any $v \neq 0 \in \mathbb{C}^p$, we apply $\operatorname{Im} b_r(z) \geq 0$ to get

$$\operatorname{Im} v^* [z \,\mathrm{Id} + b \cdot \mathring{\Sigma} + \Gamma^\mathsf{T} \operatorname{diag}_\ell(b) \Gamma] v > 0,$$

and hence z is not a root of (5.9). A similar argument holds for $z \in \mathbb{C}^-$, which establishes (a).

Note by Proposition 5.11 and Assumption 2.2 that $\|b \cdot \mathring{\Sigma} + \Gamma^\mathsf{T} \operatorname{diag}_\ell(b) \Gamma\| < C$ for large n. Then (b) follows from (5.9). For (c), note by Proposition 5.6 that

$$\|Q^\mathsf{T} R(z) Q + Q^\mathsf{T} (z \,\mathrm{Id} + b \cdot \mathring{\Sigma})^{-1} Q\| < C$$

almost surely for large n. On the event $\mathcal{E}_n$, $\|Q^\mathsf{T} R(z) Q\|$ is uniformly bounded. Then so is $\|Q^\mathsf{T} (z \,\mathrm{Id} + b \cdot \mathring{\Sigma})^{-1} Q\|$ for large n. Combining with Proposition 5.11 and Assumption 2.2, the first bound of (c) follows. Since the dimension of $\widetilde{T}(z)$ is ℓ which is at most a constant, the first bound implies the second.

We now prove the statements for $\widehat{T}(z)$. For (a'), by Lemma 5.3, $\det \widehat{K}_{11}(z)$ almost surely for large n does not vanish on U_δ. Thus, for $z \in U_\delta$, by the Schur complement formula, we see that

$$\det \widehat{K}(z) = \det \widehat{K}_{11}(z) \cdot \det \widehat{T}(z),$$

meaning that any root of $\det \widehat{T}(z) = 0$ is also a root of $\det \widehat{K}(z) = 0$, hence a (real) eigenvalue of $\widehat{\Sigma}$ by Lemma 5.2. Claim (b') follows from the fact that roots of $\det \widehat{T}(z) = 0$ are eigenvalues of $\widehat{\Sigma}$, and $\|\widehat{\Sigma}\| < C$ almost surely for large n. For (c'), we apply $\|\widetilde{T}(z)\|_\infty < C$ and $\widehat{T}(z) \sim \dot{T}(z) \sim \widetilde{T}(z)$ from Lemmas 5.3 and 5.5. □

We now establish the following technical result which will allow us to pass to convergent subsequences.

Proposition 5.14. *There exists a subsequence* $\{n_l^0\}_{l=1}^\infty$ *along which* $\operatorname{supp}(\mu_0)$ *converges to a fixed closed set* $V \subset \mathbb{R}$ *in the sense that*

$$\lim_{l \to \infty} \sup_{z \in V} \operatorname{dist}(z, \operatorname{supp}(\mu_0)) = 0 \qquad \text{and} \qquad \lim_{l \to \infty} \sup_{z \in \operatorname{supp}(\mu_0)} \operatorname{dist}(z, V) = 0, \tag{5.10}$$

and $\det \widetilde{T}(z)$ *converges to a fixed analytic function* $D : \mathbb{C} \setminus V \to \mathbb{C}$ *uniformly on compact subsets.*

Proof. The first part of the statement follows from Proposition 5.12 and sequential compactness of the metric space of compact subsets of $[-C, C]$ under the Hausdorff metric. For the second part, note that $\det \widetilde{T}(z)$ is well-defined and analytic on compact subsets of $\mathbb{C} \setminus V$ at $n = n_l^0$ for all large l. Proposition 5.13 ensures that $\det \widetilde{T}(z)$ is uniformly bounded on any such compact subset. Then Montel's Theorem implies that there is a further subsequence which converges uniformly over compact sets to an analytic function D. □

Proof of Theorem 2.6. Note that by the identity $\det(\mathrm{Id}+XY) = \det(\mathrm{Id}+YX)$, Λ_0 is also the set of roots of $0 = \det \widetilde{T}(z)$. Let Ω be the sample space, and $\Omega_0 \subset \Omega$ the event of probability 1 on which all preceding almost sure statements hold.

Fix $\omega \in \Omega_0$. First suppose that we pass to a subsequence satisfying the result of Proposition 5.14, meaning that $\mathrm{supp}(\mu_0)$ converges to a fixed closed set V and $\det \widetilde{T}(z) \to D(z)$ uniformly on compact subsets of $\mathbb{C} \setminus V$. By Proposition 5.13(a), for all large n, all roots of $\det \widetilde{T}(z) = 0$ and $\det \widehat{T}(z) = 0$ with distance at least $\delta/2$ to V are real and have magnitude less than some $R > 0$. Because $\det \widetilde{T}(z) \to D(z)$, we see that this is true for D as well. Since D is analytic, this implies that D has finitely many such roots. Let

$$\lambda_1 < \cdots < \lambda_J$$

be the distinct roots of D whose distance to V is at least $\delta/2$, and let m_j be the multiplicity of λ_j.

Choose ε small enough so that $\varepsilon < \delta/4$. For constants $r_j, \sigma > 0$, let γ_j be the counterclockwise contour traversing the rectangle with vertices $(\lambda_j \pm r_j) \pm i\sigma$. Choose r_j, σ small enough such that

- the contours γ_j do not intersect,
- each γ_j is contained within a radius $\varepsilon/2$ ball centered at λ_j, and
- the only root of $D(z)$ contained within or on each γ_j is λ_j.

Partitioning the set

$$\{x \in \mathbb{R} : \mathrm{dist}(x, \mathrm{supp}(\mu_0)) > \delta/2,\ \mathrm{dist}(x, \lambda_j) > r_j \text{ for all } j,\ |x| < R\}$$

into disjoint open intervals, for each such interval $\mathcal{I} = (l, u)$, define also a counterclockwise contour $\gamma'_{\mathcal{I}}$ traversing the rectangle with vertices $l \pm i\sigma$ and $u \pm i\sigma$.

By construction, $D(z)$ does not vanish along any of γ_j or $\gamma'_{\mathcal{I}}$, and $\det \widetilde{T}(z)$ converges uniformly to $D(z)$ on each contour. Hence, by Hurwitz's theorem, for all large n, $\det \widetilde{T}(z)$ has m_j zeros within each γ_j, which are real by Proposition 5.13, and no zeros within each $\gamma'_{\mathcal{I}}$. Now, observe that by Lemmas 5.3 and 5.5, as $n \to \infty$,

$$\sup_{z \in U_{\delta/4}} \|\widehat{T}(z) - \widetilde{T}(z)\|_\infty \to 0,$$

which implies by Proposition 5.13 that $|\det \widehat{T}(z) - \det \widetilde{T}(z)| \to 0$ uniformly on each contour and thus that $|\det \widehat{T}(z) - D(z)| \to 0$ uniformly on each contour. Applying Hurwitz's theorem again, we find that for all large n, $\det \widehat{T}(z)$ also has m_j zeros within each γ_j, which are real by Proposition 5.13, and no zeros within each $\gamma'_{\mathcal{I}}$.

Taking Λ_δ and $\hat{\Lambda}_\delta$ as the zeros of $\det \widetilde{T}(z)$ and $\det \widehat{T}(z)$ within the contours γ_j, this yields

$$\text{ordered-dist}(\Lambda_\delta, \hat{\Lambda}_\delta) < \varepsilon.$$

By Lemma 5.3, for $z \in U_\delta$, $\widehat{K}_{11}(z)$ is invertible for large n, so we may apply the Schur complement formula to obtain

$$\det \widehat{K}(z) = \det \widehat{K}_{11}(z) \det \widehat{T}(z).$$

By Lemma 5.2, we conclude that $\widehat{\Lambda}_\delta \subseteq \mathrm{spec}(\widehat{\Sigma})$. Further, since neither $\det \widetilde{T}(z)$ or $\det \widehat{T}(z)$ have zeros inside $\gamma'_{\mathcal{I}}$ or $(-\infty, R] \cup [R, \infty)$, we find that Λ_δ and $\hat{\Lambda}_\delta$ contain all zeros of $\det \widetilde{T}(z)$ and elements of $\mathrm{spec}(\widehat{\Sigma})$, respectively, which have distance at least $\delta/2$ from V. Thus they contain all such values which have distance at least δ from $\mathrm{supp}(\mu_0)$ for all large n, establishing the result along this subsequence.

To conclude the proof, suppose by contradiction that there is a subset $\Omega_1 \subset \Omega_0$ of positive probability for which there is a ω-dependent subsequence $\{n_l^0\}$ such that for each $n = n_l^0$, no such sets Λ_δ and $\hat{\Lambda}_\delta$ satisfying the required conditions exist. By Proposition 5.14, there is a further subsequence along which $\mathrm{supp}(\mu_0)$ and $\det \widetilde{T}(z)$ converge. On this subsequence, our previous construction shows that Λ_δ and $\hat{\Lambda}_\delta$ satisfying the desired conditions exist, a contradiction. This concludes the proof. □

5.6. **Outlier eigenvectors.** Finally, we prove Theorem 2.7 on the alignments of eigenvectors for isolated outliers. The proof will proceed in two steps. First, in the following result we bound the second smallest singular value of $\widetilde{T}(\lambda)$ at $\lambda \in \Lambda_\delta$. Second, we use perturbation theory to relate the eigenvector for an isolated outlier to a vector in $\ker \widetilde{T}(\lambda)$.

Proposition 5.15. *In the setting of Theorem 2.7,* $\ker \widetilde{T}(\lambda)$ *has dimension exactly* 1*, and each other singular value of* $\widetilde{T}(\lambda)$ *is at least a constant* $c \equiv c(\delta) > 0$*.*

Proof. Recall from (5.9) that the roots of $\det \widetilde{T}(z) = 0$ are also roots of $\det M(z) = 0$ for $M(z) = z\,\mathrm{Id} + b \cdot \mathring{\Sigma} + \Gamma^\mathsf{T} \operatorname{diag}_\ell(b)\Gamma = z\,\mathrm{Id} + b \cdot \Sigma$. By Proposition 5.1, we see that

$$\partial_z[z\,\mathrm{Id} + b(z) \cdot \Sigma] = \mathrm{Id} + b'(z) \cdot \Sigma \succeq \mathrm{Id}$$

for $z \in \mathbb{R} \setminus \operatorname{supp}(\mu_0)$, meaning that the ordered eigenvalues of $M(z)$ increase at a rate of at least 1 on each interval of $\mathbb{R} \setminus \operatorname{supp}(\mu_0)$. As a result, at the given isolated root $z = \lambda$ of $0 = \det M(z)$, the matrix $M(\lambda)$ has a single eigenvalue equal to 0 and remaining eigenvalues outside $(-\delta, \delta)$, and the second-smallest singular value of $M(\lambda)$ is at least δ. By Propositions 5.11 and 5.13(b), we see that $|b_r(\lambda)|$ and $|\lambda|$ are bounded, so $\|\lambda \cdot \mathrm{Id} + b(\lambda) \cdot \mathring{\Sigma}\| < C$ and all singular values of $(\lambda\,\mathrm{Id} + b(\lambda) \cdot \mathring{\Sigma})^{-1}$ are at least $1/C$. Now, letting $v \in \ker M(\lambda)$ be a unit vector, we have that for any $w \in \mathbb{R}^p$, $\|w\|^2 - |v^\mathsf{T} w|^2$ is the squared length of the component of w orthogonal to v. Then

$$\|(\lambda\,\mathrm{Id} + b \cdot \mathring{\Sigma})^{-1} M(\lambda) w\|^2 \geq (\delta/C)^2 \cdot (\|w\|^2 - |v^\mathsf{T} w|^2). \tag{5.11}$$

Choose $U \in \mathbb{R}^{p \times (p-\ell)}$ so that $[Q \mid U]$ is an orthogonal matrix. Notice that because

$$(\lambda\,\mathrm{Id} + b \cdot \mathring{\Sigma})^{-1}(\lambda\,\mathrm{Id} + b \cdot \Sigma)U = U + (\lambda\,\mathrm{Id} + b \cdot \mathring{\Sigma})^{-1}\Gamma^\mathsf{T} \operatorname{diag}_\ell(b)\Gamma U = U,$$

we have that

$$[Q \mid U]^\mathsf{T}(\lambda\,\mathrm{Id} + b \cdot \mathring{\Sigma})^{-1} M(\lambda)[Q \mid U] = \begin{bmatrix} \widetilde{T}(\lambda) & 0 \\ U^\mathsf{T}(\lambda\,\mathrm{Id} + b \cdot \mathring{\Sigma})^{-1}(\Gamma^\mathsf{T} \operatorname{diag}_\ell(b)\Gamma)Q & \mathrm{Id} \end{bmatrix}.$$

Note that $(Q^\mathsf{T} v, U^\mathsf{T} v)$ is a unit vector in the kernel of this matrix, hence $Q^\mathsf{T} v \in \ker \widetilde{T}(\lambda)$ and $Q^\mathsf{T} v \neq 0$. Now, for any $u_1 \in \mathbb{R}^\ell$ orthogonal to $Q^\mathsf{T} v$ and

$$u_2 = -U^\mathsf{T}(\lambda\,\mathrm{Id} + b \cdot \mathring{\Sigma})^{-1}(\Gamma^\mathsf{T} \operatorname{diag}_\ell(b)\Gamma)Q u_1,$$

define the vector $u = (u_1, u_2)$. For this u, we obtain by (5.11) that

$$\|\widetilde{T}(\lambda)u_1\|^2 = \|[Q \mid U]^\mathsf{T}(\lambda\,\mathrm{Id} + b \cdot \mathring{\Sigma})^{-1} M(\lambda)[Q \mid U]u\|^2 \geq (\delta/C)^2(\|u\|^2 - |v^\mathsf{T} Q u_1 + v^\mathsf{T} U u_2|^2).$$

Since u_1 is orthogonal to $Q^\mathsf{T} v$ and v is a unit vector, we see that

$$|v^\mathsf{T} Q u_1 + v^\mathsf{T} U u_2| = |v^\mathsf{T} U u_2| \leq \|u_2\|.$$

Substituting, we obtain for any u_1 orthogonal to $Q^\mathsf{T} v$ that $\|\widetilde{T}(\lambda)u_1\|^2 \geq (\delta/C)^2 \|u_1\|^2$. This implies that $\ker \widetilde{T}(\lambda)$ is one-dimensional and spanned by $Q^\mathsf{T} v$, and the next smallest singular value of $\widetilde{T}(\lambda)$ is bounded below by δ/C, as desired. □

Proposition 5.16. *Denote by* $S'(z)$ *and* $R'(z)$ *the derivatives of* $S(z)$ *and* $R(z)$ *with respect to* z*. Then*

$$S'(z) \sim 0, \qquad Q^\mathsf{T} R'(z) Q \sim -Q^\mathsf{T} \partial_z[(z\,\mathrm{Id} + b \cdot \mathring{\Sigma})^{-1}]Q, \qquad n_t^{-1} \operatorname{Tr}_t[FGHR'(z)H^\mathsf{T} G^\mathsf{T} F] \sim b_t'(z).$$

Proof. By Lemma 5.3, we have

$$\sup_{z \in U_{\delta/2}} \|S(z)\|_\infty \to 0$$

almost surely. For each $z \in U_\delta$, define a contour $\gamma(t) = \delta/2 \cdot e^{it}$ for $t \in [0, 2\pi]$. Applying the Cauchy integral formula entrywise to $S(z)$, we get

$$\|S'(z)\|_\infty \leq \frac{2}{\delta} \cdot \max_{t \in [0,2\pi]} \|S(z + \gamma(t))\|_\infty \leq C \sup_{z \in U_{\delta/2}} \|S(z)\|_\infty.$$

Hence $S'(z) \sim 0$. The other statements follow similarly from Propositions 5.6 and 5.7. □

Proof of Theorem 2.7. Since $(\widehat{\lambda}, \widehat{v})$ is an eigenvalue-eigenvector pair, we have that $\widehat{\lambda}\widehat{v} = \widehat{\Sigma}\widehat{v} = W\widehat{v} + P\widehat{v}$, which implies that

$$0 = (\mathrm{Id} + R(\widehat{\lambda})P)\widehat{v}. \tag{5.12}$$

Define

$$\widehat{v}_1 = \Xi^\mathsf{T} FGH\widehat{v} \qquad \text{and} \qquad \widehat{v}_2 = Q^\mathsf{T}\widehat{v}.$$

Multiplying the previous equation on the left by $\begin{bmatrix}\Xi^\mathsf{T} FGH \\ Q^\mathsf{T}\end{bmatrix}$, we find that

$$0 = \Big(\mathrm{Id} + \begin{bmatrix}\Xi^\mathsf{T} FGH \\ Q^\mathsf{T}\end{bmatrix} R(\widehat{\lambda}) \begin{bmatrix} Q\widetilde{\Gamma}^\mathsf{T} & H^\mathsf{T} G^\mathsf{T} F \Xi \widetilde{\Gamma} + Q\widetilde{\Gamma}^\mathsf{T} \Xi^\mathsf{T} F \Xi \widetilde{\Gamma}\end{bmatrix}\Big) \begin{bmatrix}\widehat{v}_1 \\ \widehat{v}_2\end{bmatrix} = \widehat{K}(\widehat{\lambda}) \begin{bmatrix}\widehat{v}_1 \\ \widehat{v}_2\end{bmatrix}.$$

Eliminating $\widehat{v}_1$ in this system of equations by taking the Schur complement, we obtain $0 = \widehat{T}(\widehat{\lambda})\widehat{v}_2$. On the event $\mathcal{E}_n$ of (5.5), $\widetilde{T}'(z)$ is bounded over U_δ. Then by the implications $\widehat{\lambda} - \lambda \to 0$ and $\widehat{T} \sim \widetilde{T}$ of Theorem 2.6 and Lemmas 5.3 and 5.5, almost surely $\|\widehat{T}(\widehat{\lambda}) - \widetilde{T}(\lambda)\| \to 0$. Then also

$$\|\widehat{T}(\widehat{\lambda})^\mathsf{T} \widehat{T}(\widehat{\lambda}) - \widetilde{T}(\lambda)^\mathsf{T} \widetilde{T}(\lambda)\| \to 0.$$

Applying this to $\widehat{v}_2$, we find that

$$\|\widetilde{T}(\lambda)^\mathsf{T} \widetilde{T}(\lambda)\widehat{v}_2\| \to 0,$$

which implies by Proposition 5.15 and the Davis-Kahan theorem that

$$\widehat{v}_2 - \|\widehat{v}_2\| v_2 \to 0, \tag{5.13}$$

where v_2 is a unit vector in $\ker \widetilde{T}(\lambda)$ with an appropriate choice of sign.

We now compute the limit of $\|\widehat{v}_2\|$. By (5.12) and the definition of P, we see that

$$-\widehat{v} = R(\widehat{\lambda})\Big(Q\widetilde{\Gamma}^\mathsf{T}\widehat{v}_1 + (H^\mathsf{T} G^\mathsf{T} F \Xi \widetilde{\Gamma} + Q\widetilde{\Gamma}^\mathsf{T}\Xi^\mathsf{T} F \Xi \widetilde{\Gamma})\widehat{v}_2\Big). \tag{5.14}$$

On the other hand, in the equation $0 = \widehat{K}(\widehat{\lambda}) \begin{bmatrix}\widehat{v}_1 \\ \widehat{v}_2\end{bmatrix}$, we may solve for $\widehat{v}_1$ to obtain

$$\widehat{v}_1 = -\widehat{K}_{11}(\widehat{\lambda})^{-1} \widehat{K}_{12}(\widehat{\lambda})\widehat{v}_2$$

when $\widehat{K}_{11}(\widehat{\lambda})$ is invertible. Substituting into (5.14), we obtain

$$\widehat{v} = R(\widehat{\lambda})(M_1(\widehat{\lambda}) + M_2(\widehat{\lambda}))\widehat{v}_2 \tag{5.15}$$

for the matrices

$$M_1(\widehat{\lambda}) = Q\widetilde{\Gamma}^\mathsf{T} \widehat{K}_{11}(\widehat{\lambda})^{-1} \widehat{K}_{12}(\widehat{\lambda}) - Q\widetilde{\Gamma}^\mathsf{T}\Xi^\mathsf{T} F \Xi \widetilde{\Gamma}, \quad M_2(\widehat{\lambda}) = -H^\mathsf{T} G^\mathsf{T} F \Xi \widetilde{\Gamma}.$$

Taking the norm of (5.15) on both sides and applying again $\widehat{\lambda} - \lambda \to 0$ and a derivative bound on $\mathcal{E}_n$,

$$1 = \sum_{i,j=1}^{2} \widehat{v}_2^\mathsf{T} M_i(\widehat{\lambda})^\mathsf{T} R(\widehat{\lambda})^2 M_j(\widehat{\lambda})\widehat{v}_2 = \sum_{i,j=1}^{2} \widehat{v}_2^\mathsf{T} M_i(\lambda)^\mathsf{T} R(\lambda)^2 M_j(\lambda)\widehat{v}_2 + o(1). \tag{5.16}$$

Applying Lemma 5.3 and Propositions 5.4 and 5.7, we find that

$$Q^\mathsf{T} M_1(z) \sim \widetilde{\Gamma}^\mathsf{T}\Xi^\mathsf{T} FGHR(z)H^\mathsf{T} G^\mathsf{T} F\Xi\widetilde{\Gamma} - \widetilde{\Gamma}^\mathsf{T}\Xi^\mathsf{T} F\Xi\widetilde{\Gamma} \sim \widetilde{\Gamma}^\mathsf{T} \operatorname{diag}_\ell(b(z))\widetilde{\Gamma}.$$

Also, noting that $R(z)^2 = R'(z)$ and applying Proposition 5.16,

$$Q^\mathsf{T} R(z)^2 Q \sim Q^\mathsf{T} R'(z) Q \sim -Q^\mathsf{T} \partial_z[(z\,\mathrm{Id} + b(z) \cdot \mathring{\Sigma})^{-1}]Q.$$

Combining these, applying $\Gamma = \widetilde{\Gamma} Q^\mathsf{T}$, and setting

$$\widehat{u} = \widetilde{\Gamma}\widehat{v}_2 = \Gamma\widehat{v},$$

we get

$$\widehat{v}_2^\mathsf{T} M_1(\lambda)^\mathsf{T} R(\lambda)^2 M_1(\lambda)\widehat{v}_2 = -\widehat{u}^\mathsf{T} \operatorname{diag}_\ell(b)\Gamma \cdot \partial_\lambda[(\lambda\,\mathrm{Id} + b \cdot \mathring{\Sigma})^{-1}] \cdot \Gamma^\mathsf{T} \operatorname{diag}_\ell(b)\widehat{u} + o(1) \tag{5.17}$$

where we write as shorthand $b \equiv b(\lambda)$. Applying $R(z)^2 = R'(z)$ and Propositions 5.4 and 5.16, we also get

$$\Xi^\mathsf{T} FGHR(z)^2 H^\mathsf{T} G^\mathsf{T} F\Xi \sim \operatorname{diag}_\ell(b'(z)),$$

and hence

$$\widehat{v}_2^\mathsf{T} M_2(\lambda)^\mathsf{T} R(\lambda)^2 M_2(\lambda)\widehat{v}_2 = \widehat{u}^\mathsf{T} \operatorname{diag}_\ell(b')\widehat{u} + o(1). \tag{5.18}$$

Finally, applying $S'(z) \sim 0$ from Proposition 5.16, we get $Q^\mathsf{T} R(z)^2 H^\mathsf{T} G^\mathsf{T} F\Xi \sim 0$ and hence

$$\widehat{v}_2^\mathsf{T} M_1(\lambda)^\mathsf{T} R(\lambda)^2 M_2(\lambda)\widehat{v}_2 \to 0. \tag{5.19}$$

Then substituting (5.17), (5.18), and (5.19) into (5.16),

$$1 = \widehat{u}^\mathsf{T}\Big(-\operatorname{diag}_\ell(b)\Gamma \cdot \partial_\lambda[(\lambda\,\mathrm{Id} + b \cdot \mathring{\Sigma})^{-1}] \cdot \Gamma^\mathsf{T} \operatorname{diag}_\ell(b) + \operatorname{diag}_\ell(b')\Big)\widehat{u} + o(1). \tag{5.20}$$

Multiplying (5.13) on the left by $\widetilde{\Gamma}$, we find that

$$\widehat{u} - \|\widehat{v}_2\|\widetilde{\Gamma}v_2 \to 0. \tag{5.21}$$

Define $\tilde{u} = \widetilde{\Gamma}v_2$, and note that $\tilde{u}$ is a non-zero vector in $\ker T(\lambda)$ because v_2 is a unit vector in $\ker \widetilde{T}(\lambda)$. Then $u = \tilde{u}/\|\tilde{u}\|$ is a unit vector in $\ker T(\lambda)$, which is unique up to sign by Proposition 5.15. Substituting (5.21) into (5.20) and recalling the definition of α in Theorem 2.7, we find that

$$1 = \|\widehat{v}_2\|^2\|\tilde{u}\|^2 \cdot \alpha + o(1).$$

Writing (5.21) as $\widehat{u} - \|\widehat{v}_2\|\|\tilde{u}\|u \to 0$ and substituting $\alpha^{-1/2}$ for $\|\widehat{v}_2\|\|\tilde{u}\|$ concludes the proof. □

References

[AGZ10] Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. *An Introduction to Random Matrices*. Cambridge University Press, Cambridge New York, 2010.

[BB16a] Monika Bhattacharjee and Arup Bose. Large sample behaviour of high dimensional autocovariance matrices. *The Annals of Statistics*, 44(2):598–628, 2016.

[BB16b] Monika Bhattacharjee and Arup Bose. Polynomial generalizations of the sample variance-covariance matrix when $pn^{-1} \to 0$. *Random Matrices: Theory and Applications*, 5(04):1650014, 2016.

[BB17] Monika Bhattacharjee and Arup Bose. Matrix polynomial generalizations of the sample variance-covariance matrix when $pn^{-1} \to y \in (0, \infty)$. *Indian Journal of Pure and Applied Mathematics*, 48(4):575–607, 2017.

[BBC17] Serban Belinschi, Hari Bercovici, and Mireille Capitaine. On the outlying eigenvalues of a polynomial in large independent random matrices. *arXiv preprint arXiv:1703.08102*, 2017.

[BBCF17] Serban T Belinschi, Hari Bercovici, Mireille Capitaine, and Maxime Février. Outliers in the spectrum of large deformed unitarily invariant models. *The Annals of Probability*, 45(6A):3571–3625, 2017.

[BBP05] Jinho Baik, Gerard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[BC17] Serban T Belinschi and Mireille Capitaine. Spectral properties of polynomials in independent Wigner and deterministic matrices. *Journal of Functional Analysis*, 273(12):3901–3963, 2017.

[BEK+14] Alex Bloemendal, László Erdős, Antti Knowles, Horng-Tzer Yau, and Jun Yin. Isotropic local laws for sample covariance and generalized Wigner matrices. *Electronic Journal of Probability*, 19(33):1–53, 2014.

[BG09] Florent Benaych-Georges. Rectangular random matrices, related convolution. *Probability Theory and Related Fields*, 144(3-4):471–515, 2009.

[BGN11] Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

[BGN12] Florent Benaych-Georges and Raj Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.

[BJW05] Zdzisław Burda, Jerzy Jurkiewicz, and Bartłomiej Wacław. Spectral moments of correlated Wishart matrices. *Physical Review E*, 71(2):026111, 2005.

[BM15] Mark W Blows and Katrina McGuigan. The distribution of genetic variance across phenotypic space and the response to selection. *Molecular Ecology*, 24(9):2056–2072, 2015.

[BS06] Jinho Baik and Jack W Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408, 2006.

[BY12] Zhidong Bai and Jianfeng Yao. On sample eigenvalues in a generalized spiked population model. *Journal of Multivariate Analysis*, 106:167–177, 2012.

[CC04] Mireille Capitaine and Muriel Casalis. Asymptotic freeness by generalized moments for Gaussian and Wishart matrices. Application to Beta random matrices. *Indiana University mathematics journal*, 53(2):397–431, 2004.

[CDM07] Mireille Capitaine and Catherine Donati-Martin. Strong asymptotic freeness for Wigner and Wishart matrices. *Indiana University mathematics journal*, 56(2):767–803, 2007.

[CHS18] Benoit Collins, Takahiro Hasebe, and Noriyoshi Sakuma. Free probability for purely discrete eigenvalues of random matrices. *Journal of the Mathematical Society of Japan*, 70(3):1111–1150, 2018.

[CMA+18] Julie M Collet, Katrina McGuigan, Scott L Allen, Stephen F Chenoweth, and Mark W Blows. Mutational pleiotropy and the strength of stabilizing selection within and between functional modules of gene expression. *Genetics*, 208(4):1601–1616, 2018.

[Col03] Benoit Collins. Moments and cumulants of polynomial random variables on unitary groups, the Itzykson-Zuber integral, and free probability. *International Mathematics Research Notices*, 2003(17):953–982, 2003.

[CŚ06] Benoit Collins and Piotr Śniady. Integration with respect to the Haar measure on unitary, orthogonal and symplectic group. *Communications in Mathematical Physics*, 264(3):773–795, 2006.

[DL18] Edgar Dobriban and Sifan Liu. A new theory for sketching in linear regression. *arXiv preprint arXiv:1810.06089*, 2018.

[Dyk93] Ken Dykema. On certain free product factors via an extended matrix model. *Journal of Functional Analysis*, 112(1):31–60, 1993.

[Fis18] Ronald A Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(02):399–433, 1918.

[FJ16] Zhou Fan and Iain M Johnstone. Eigenvalue distributions of variance components estimators in high-dimensional random effects models. *arXiv preprint arXiv:1607.02201*, 2016.

[FJ17] Zhou Fan and Iain M Johnstone. Tracy-Widom at each edge of real covariance estimators. *arXiv preprint arXiv:1707.02352v2*, 2017.

[FJS18] Zhou Fan, Iain M Johnstone, and Yi Sun. Spiked covariances and principal components analysis in high-dimensional random effects models. *arXiv preprint arXiv:1806.09529*, 2018.

[GH03] Jeffrey S Geronimo and Theodore P Hill. Necessary and sufficient condition that the limit of Stieltjes transforms is a Stieltjes transform. *Journal of Approximation Theory*, 121(1):54–60, 2003.

[HB06] Emma Hine and Mark W Blows. Determining the effective dimensionality of the genetic variance–covariance matrix. *Genetics*, 173(2):1135–1144, 2006.

[HLN07] Walid Hachem, Philippe Loubaton, and Jamal Najim. Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3):875–930, 2007.

[HP00] Fumio Hiai and Denes Petz. Asymptotic freeness almost everywhere for random matrices. *Acta Sci. Math. (Szeged)*, 66(3–4):809–834, 2000.

[HT05] Uffe Haagerup and Steen Thorbjørnsen. A new application of random matrices: $\mathrm{Ext}(c^*_{\mathrm{red}}(f_2))$ is not a group. *Annals of Mathematics*, 162:711–775, 2005.

[JL09] Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

[Joh01] Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

[Jol11] Ian Jolliffe. *Principal component analysis*. Springer, 2011.

[JP18] Iain M Johnstone and Debashis Paul. PCA in high dimensions: An orientation. *Proceedings of the IEEE*, 106(8):1277–1292, 2018.

[KY17] Antti Knowles and Jun Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169(1-2):257–352, 2017.

[LAP15] Haoyang Liu, Alexander Aue, and Debashis Paul. On the Marčenko–Pastur law for linear time series. *The Annals of Statistics*, 43(2):675–712, 2015.

[LS07] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161, 2007.

[LW98] Michael Lynch and Bruce Walsh. *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA, 1998.

[Mal12] Camille Male. The norm of polynomials in large random and deterministic matrices. *Probability Theory and Related Fields*, 154(3-4):477–532, 2012.

[MP67] Vladimir A Marcenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Sbornik: Mathematics*, 1(4):457–483, 1967.

[MS17] James A Mingo and Roland Speicher. *Free probability and random matrices*, volume 35. Springer, 2017.

[Nad08] Boaz Nadler. Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, 36(6):2791–2817, 2008.

[NS06] Alexandru Nica and Roland Speicher. *Lectures on the Combinatorics of Free Probability*. Cambridge University Press, 2006.

[NSS02] Alexandru Nica, Dimitri Shlyakhtenko, and Roland Speicher. Operator-valued distributions. I. Characterizations of freeness. *International Mathematics Research Notices*, 29:1509–1538, 2002.

[PA14] Debashis Paul and Alexander Aue. Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150:1–29, 2014.

[Pau07] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642, 2007.

[Rao72] C Radhakrishna Rao. Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association*, 67(337):112–115, 1972.

[Sch05] Hanne Schultz. Non-commutative polynomials of independent Gaussian random matrices. The real and symplectic cases. *Probability Theory and Related Fields*, 131(2):261–309, 2005.

[SCM09] Shayle R Searle, George Casella, and Charles E McCulloch. *Variance Components*. John Wiley & Sons, 2009.

[Shl15] Dimitri Shlyakhtenko. Free probability of type B and asymptotics of finite-rank perturbations of random matrices. *arXiv preprint arXiv:1509.08841*, 2015.

[SPP+12] Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*, 7(3):500, 2012.

[SV12] Roland Speicher and Carlos Vargas. Free deterministic equivalents, rectangular random matrix models, and operator-valued free probability theory. *Random Matrices: Theory and Applications*, 1(02):1150008, 2012.

[TW96] Craig A Tracy and Harold Widom. On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics*, 177(3):727–754, 1996.

[Voi91] Dan Voiculescu. Limit laws for random matrices and free products. *Inventiones mathematicae*, 104(1):201–220, 1991.

[Voi98] Dan Voiculescu. A strengthened asymptotic freeness result for random matrices with applications to free entropy. *International Mathematics Research Notices*, 1998(1):41–63, 1998.

[WAP17] Lili Wang, Alexander Aue, and Debashis Paul. Spectral analysis of sample autocovariance matrices of a class of linear time series in moderately high dimensions. *Bernoulli*, 23(4A):2181–2209, 2017.

[Wri35] Sewall Wright. The analysis of variance and the correlations between relatives with respect to deviations from an optimum. *Journal of Genetics*, 30(2):243–256, 1935.

[YLGV11] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

[Zha06] Lixin Zhang. *Spectral analysis of large dimensional random matrices*. PhD thesis, National University of Singapore, 2006.

[ZS12] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821, 2012.