X. Ding and T. Trogdon (2024) “A Riemann–Hilbert Approach to the Perturbation Theory for Orthogonal Polynomials: Applications to Numerical Linear Algebra and Random Matrix Theory,”
International Mathematics Research Notices, Vol. 2024, No. 5, pp. 3975–4061
Advance Access Publication July 4, 2023
https://doi.org/10.1093/imrn/rnad142

A Riemann–Hilbert Approach to the Perturbation Theory for Orthogonal Polynomials: Applications to Numerical Linear Algebra and Random Matrix Theory

Xiucai Ding[1] and Thomas Trogdon[2,*]

[1]Department of Statistics, University of California, Davis, Davis, CA 95616, USA and [2]Department of Applied Mathematics, University of Washington, Seattle, WA 98195, USA

[*]*Correspondence to be sent to: e-mail: trogdon@uw.edu*

We establish a new perturbation theory for orthogonal polynomials using a Riemann–Hilbert approach and consider applications in numerical linear algebra and random matrix theory. This new approach shows that the orthogonal polynomials with respect to two measures can be effectively compared using the difference of their Stieltjes transforms on a suitably chosen contour. Moreover, when two measures are close and satisfy some regularity conditions, we use the theta functions of a hyperelliptic Riemann surface to derive explicit and accurate expansion formulae for the perturbed orthogonal polynomials. In contrast to other approaches, a key strength of the methodology is that estimates can remain valid as the degree of the polynomial grows. The results are applied to analyze several numerical algorithms from linear algebra, including the Lanczos tridiagonalization procedure, the Cholesky factorization, and the conjugate gradient algorithm. As a case study, we investigate these algorithms applied to a general spiked sample covariance matrix model by considering the eigenvector empirical spectral distribution and its limits. For the first time, we give precise estimates on the output of the algorithms, applied to this wide class of random matrices, as the number of iterations diverges. In this setting, beyond the first order expansion, we also derive a new mesoscopic central limit theorem for the associated orthogonal polynomials and other quantities relevant to numerical algorithms.

Received November 30, 2022; Revised May 16, 2023; Accepted June 8, 2023
Communicated by Prof. Ivan Corwin

1 Introduction

We consider a Riemann–Hilbert approach to the perturbation of orthogonal polynomials. More specifically, we present an approach to compare the orthogonal polynomials with respect to two compactly supported measures on $\mathbb{R}$ by comparing their Stieltjes transforms on a contour that encircles and contracts to the union of the supports. The approach uses and generalizes the Fokas–Its–Kitaev reformulation of orthogonal polynomials [41] as the solution of a Riemann–Hilbert problem. This approach is especially powerful when the orthogonal polynomials with respect to one of the measures has known asymptotics. And in particular, it allows one to compare, in a convenient framework, polynomials orthogonal to a discrete empirical measure, that is, discrete orthogonal polynomials, to the polynomials orthogonal with respect to a limiting measure. We refer the reader to [4] for many related details concerning discrete orthogonal polynomials.

Measures are often compared rather effectively using their moments. But even measures that are rather close in a variety of senses can have vastly different moments of high order. For this reason, many studies of the perturbations of orthogonal polynomials are not infinitesimal in nature, see [42] and the references therein, particularly [78]. One construction of orthogonal polynomials uses their representation in terms of determinants of Hankel moment matrices (see [18] and [43], for example). This fact was recently exploited in [33, 64] to compare two sequences of orthogonal polynomials when the degree is bounded. But as the degree increases, this approach fails because two sequences of orthogonal polynomials with respect to two similar measures typically deviate exponentially, see [43, Section 2.1.6]. But the Fokas–Its–Kitaev Riemann–Hilbert problem gives a mechanism to make sense of the behavior of one sequence of orthogonal polynomials relative to another, giving a sense in which the mapping from a Stieltjes transform of a measure to the associated orthogonal polynomials (and their weighted Cauchy integrals) is well conditioned.

Comparing sequences of orthogonal polynomials via their Stieltjes transforms lends itself directly to estimates from random matrix theory. For example, the well-known local laws for Wigner, generalized Wigner, and (spiked) sample covariance matrices are precisely comparisons of Stieltjes transforms of measures on contours approaching the supports on small scales; see the monograph [39] for more details. Importantly, the standard empirical spectral distributions associated with these matrices, measures that weight each eigenvalue equally, are not as likely to arise in applications from computational mathematics. So one, in turn, looks to the so-called anisotropic local laws [51], which gives, in particular, the comparison of the Stieltjes transform of the eigenvector empirical spectral distribution (VESD), which, for an $N \times N$ symmetric matrix W and vector $\mathbf{b}$, is given by [2],

$$\nu = \sum_{j=1}^{N} |\langle \mathbf{q}_j, \mathbf{b} \rangle|^2 \delta_{\lambda_j(W)}, \tag{1.1}$$

where $\mathbf{q}_j$ is a normalized eigenvector associated with eigenvalue $\lambda_j(W)$ of W. For the sake of completeness, we note that if the weights $|\langle \mathbf{q}_j, \mathbf{b} \rangle|^2$ are each replaced with $1/N$ the resulting measure is called the empirical spectral distribution (ESD).

Our main application of the estimates for random polynomials orthogonal to the VESD concerns the (bi/tri)diagonalization of random matrices and, as a consequence, applications to other critically important numerical algorithms acting on random matrices, see Section 3.1 for more details. Here, we take the tridiagonalization as an example. Going back to the work of Silverstein [69], and the subsequent work of Dumitriu and Edelman [37], it is well-known that the tridiagonalization T of a Wishart matrix $W = XX^*$, where $X_{ij} \overset{\mathscr{L}}{=} \mathcal{N}(0, M^{-1})$, and X is $N \times M$, and has independent entries, has an explicit distributional description in terms of independent χ-distributed random variables (see (5.5) below). But this description is actually derived first from a distributional description of the Cholesky decomposition (We discuss tridiagonalization and the Cholesky decomposition in Section 3.1 below.)

$$T = LL^*, \quad L = (\ell_{i,j}). \tag{1.2}$$

The Cholesky factorization in this context is a lower-bidiagonal factorization of the tridiagonalization. An immediate consequence of this bidiagonalization is that $\ell_{n,n} - \sqrt{\frac{M-n+1}{M}}$ and $\ell_{n+1,n} - \sqrt{\frac{N-n}{M}}$ tend to zero and have Gaussian fluctuations provided $M - n$ and $N - n$, respectively, tend to ∞. It is therefore natural to ask if this behavior persists for both non-Gaussian entries (universality) and if it persists for sample covariance matrices with non-trivial covariance. It was recently proved in [64] that for non-Gaussian entries with trivial covariance, if $N/M \to c \in (0, 1]$ and n is fixed one sees that the upper-left $n \times n$ subblock of L tends to the Cholesky factorization of the three-term recurrence Jacobi matrix for the orthogonal polynomials with respect to the Marchenko–Pastur law with parameter c. These arguments do not apply if either n diverges and the entries X_{ij} are non-Gaussian or if the covariance is non-trivial. Our Riemann–Hilbert approach extends these results, and the results of [33], to non-trivial covariance and unbounded n.

We summarize related results in Sections 1.1 and 1.2 and provide an overview of our results and key innovations in Section 1.3.

1.1 A new application of Riemann–Hilbert analysis in random matrix theory

In this section, we summarize some related results on the Riemann–Hilbert approach to orthogonal polynomials and various related applications and demonstrate how our approach differs. It is known from the celebrated work of Fokas, Its, and Kitaev [41] that orthogonal polynomials can be characterized as the solution of a 2×2 matrix Riemann–Hilbert problem with jump on the real line. Later on, a remarkable steepest descent method was proposed by Deift and Zhou in [24] to study the asymptotics of the modified Korteweg–de Vries equation. Since then, various extensions have been made, including to the asymptotics of orthogonal polynomials. More specifically, the extension on the unit circle was studied in [5], general measures and universality were studied in [10, 17, 21, 30, 53, 55], the biorthogonal polynomial problem was studied in [8, 49, 54, 75], and multiple orthogonal polynomials were studied in [74]. For a more comprehensive review, we refer the reader to [4, 12, 18, 25, 57]. Of particular relevance is the monograph [4]. In a slightly different form, this text contains the transformation (2.7) and the hyperelliptic Riemann surface theory employed in Appendix A.

Classically, the way in which Riemann–Hilbert problems and orthogonal polynomial theory connect to random matrix theory is very different from the framework we propose here. More precisely, Riemann–Hilbert problems historically enter random matrix theory via the analysis of orthogonal polynomials because the eigenvalues of many random matrix ensembles can be viewed as a determinantal point processes and the correlation functions have a determinantal kernel function that can be expressed as a sum of orthogonal polynomials. Consequently, using the Christoffel–Darboux formula, the eigenvalue correlation functions can be expressed in terms of the solution of a Riemann–Hilbert problem; see [18, 61] for a review. On the other hand, the gap probabilities can be represented as a Fredholm determinant and the limiting expressions themselves can be expressed in terms of the solution of a Riemann–Hilbert problem; see the monographs [47, 52] for a review. This approach, combined with the steepest descent method, allows for the large N asymptotics to be determined explicitly for various random matrix models leading to the determination of explicit limiting kernels. For example, for the Gaussian Unitary Ensemble (GUE), the correlation function for the bulk eigenvalues converge to the sine kernel [30, 38, 61] and the large gap probability of the edge eigenvalues converge to the Airy kernel [71]. We refer the readers to [12, 18, 57] for a more exhaustive discussion. The methodology has also been applied to various other random matrix models, see [9, 11, 15, 19, 26, 36, 56, 58, 75], to name but a few.

In the current paper, we do not study orthogonal polynomials and random matrices by following the classic research line above. In contrast, we apply a Riemann–Hilbert approach to study the behavior of orthogonal polynomials with respect to perturbations of the orthogonality measure. We then apply the theory to polynomials orthogonal with respect to the VESD (1.1) when W is random. The perturbations we consider are quantified by the closeness of their Stieltjes transforms. Such a setting is general. A wide class of (random) measures that can be thought of as appropriate perturbations of a deterministic measure are measures arising from widely studied random matrix models, where the local laws [39] guarantee the closeness of the limiting and empirical measures. Our new approach, and its generality, can best be summarized by the fact that while some random matrix ensembles have eigenvalue statistics that can be analyzed by orthogonal polynomial theory, all random matrices generate measures (again, see (1.1)), and the analyses of the orthogonal polynomials with respect to such a measure are important. We show exactly how this analysis can be accomplished using Riemann–Hilbert analysis.

1.2 Some related work on numerical algorithms

Our motivation comes from the analysis of various iterative numerical algorithms in linear algebra (see Section 3.1 for a review), especially when the inputs are random matrices. A common feature for these algorithms is that their analysis can be reduced to understanding certain (discrete) orthogonal polynomials and their associated Cauchy transforms (see (B.12), (B.14), (B.15), and (B.16) for illustrations). By establishing a perturbation theory for orthogonal polynomials, we are able to provide the first-order limits and asymptotic distributions (We determine distributions when the inputs are random.) related to these algorithms.

In the literature, various numerical algorithms have been studied when the inputs are random matrices. The tridiagonalization of Wishart matrix (i.e., sample covariance matrix with standard Gaussian entries) has been analyzed in [37, 69], the finite iterations of CGA for a sample covariance matrix with trivial covariance was analyzed in [29, 64], and the Toda algorithm on Wishart matrices was analyzed in [27, 28]. These analyses rely on either a Gaussian assumption or the trivial covariance assumption. The finite iterations of CGA with general covariance structure was analyzed in [33]. The general phenomenon that some algorithms have, in an appropriate sense, high concentration in their outputs even when the inputs are random data can be seen in each of these works. And quite often the performance of the algorithms under consideration is universal. We refer to the readers to [23, 31, 66, 67] for further discussions.

There has also been significant developments in the area of smoothed analysis of algorithms [68, 70]. More closely related to the current work is [62]. We leave the problem of using the current results in this context as future work.

1.3 An overview of main results

Given a probability measure μ with finite moments, we apply the Gram–Schmidt orthogonalization process to the monomials $\{1, \lambda, \lambda^2, \cdots\}$ to obtain the monic orthogonal polynomials $\pi_n(\lambda; \mu)$, $n = 0, 1, 2, \ldots$, which can be defined by

$$\pi_n(\lambda; \mu) = \lambda^n + O(\lambda^{n-1}), \quad \lambda \to \infty, \quad \int_{\mathbb{R}} \pi_n(\lambda; \mu)\pi_m(\lambda; \mu)\mu(\mathrm{d}\lambda) = 0, \quad n \neq m. \tag{1.3}$$

Given two measures μ and ν, where ν can be regarded as a perturbed or empirical version of μ, we aim to study how $\pi_n(\lambda; \mu)$ and $\pi_n(\lambda; \nu)$ relate asymptotically, both as n increases and as $\nu \to \mu$.

The starting point of our analysis is the quantity $X_n(z; \mu, \nu)$ introduced in (2.11). The motivation to use $X_n(z; \mu, \nu)$ is threefold. First, it naturally connects $\pi_n(\lambda; \mu)$ and $\pi_n(\lambda; \nu)$ and their associated Cauchy transforms (cf. (2.1)). Second, X_n is the solution of a matrix Riemann–Hilbert problem that can be explicitly formulated using the Fokas–Its–Kitaev approach. Third, the relevant quantities associated to the numerical algorithms we consider can be expressed in terms of the entries of $X_n(z; \mu, \nu)$. The Riemann–Hilbert problem for $X_n(z; \mu, \nu)$ can be solved asymptotically, and this result is recorded in Proposition 2.1. Equivalently, it establishes a new perturbation result for orthogonal polynomials. Heuristically, it states that for two compactly supported measures μ, ν on $\mathbb{R}$ such that

$$\int \frac{\nu(\mathrm{d}\lambda) - \mu(\mathrm{d}\lambda)}{\lambda - z}, \tag{1.4}$$

is sufficiently small on a contour that encircles, and is sufficiently close to $\mathrm{supp}(\mu) \cup \mathrm{supp}(\nu)$, one has for the monic polynomials π_n,

$$\pi_n(z; \nu) = \pi_n(z; \mu)(1 + f_1(z; \mu, \nu)) + f_2(z; \mu, \nu)\pi_{n-1}(z, \mu)\frac{\mathfrak{c}^{2(p-n)}}{\|\pi_{n-1}(\cdot; \mu)\|^2_{L^2(\mu)}}, \tag{1.5}$$

for functions $f_1, f_2 = \mathrm{o}(1)$ depending on the size of (1.4) and some constant $\mathfrak{c}$. Here p is the number of spikes (i.e., point masses, see cf. (2.16)); see (2.15) for more details. A further expansion of the functions f_1, f_2 determine the next order correction, which we, in view of our primary application to random matrices, call the fluctuation term.

Then, assuming that μ satisfies some regularity conditions (cf. Assumption 1), we first derive some accurate and uniformly valid asymptotic formulae for the unperturbed orthogonal polynomials utilizing theta functions on a hyperelliptic Riemann surface (cf. (A.1)). The results are stated in Theorem 2.2. By controlling a key auxiliary quantity (cf. (2.13)) in Lemma 2.3, we are able use Proposition 2.1 and Theorem 2.2 to provide asymptotic formulae for the perturbed orthogonal polynomials and their Cauchy transforms as in Theorem 2.4 and Remark 2.4. These formulae give explicitly how some critical exponential prefactors are arranged. Moreover, the leading error terms can be fully characterized by a variant of (1.4). Thus, the calculation of the fluctuations of $\pi_n(z;\nu)$ reduces to the analysis of (1.4).

We mention several points related to random matrix theory here. First, Assumption 1 is satisfied by the limiting eigenvalue or eigenvector empirical spectral distributions of many classically studied random matrix models. In this context, ν can be the eigenvalue or eigenvector empirical spectral distribution. Second, the degree n is allowed to be unbounded (with respect to some divergent parameter) and it depends on the closeness of the Stieltjes transforms of the measures μ and ν. For example, in the random matrix model setting regarding an $N \times N$ matrix, as will be discussed in Remark 2.3, n can be as large as $\mathrm{O}(N^{1/4-\epsilon})$, for some arbitrarily small constant $\epsilon > 0$ for ESD, and $\mathrm{O}(N^{1/6-\epsilon})$ for VESD. To our best knowledge, this is the first such asymptotic result allowing n to diverge.

Motivated by several important applications in numerical linear algebra, we apply Theorems 2.2 and 2.4 to analyze iterative numerical algorithms, including Lanczos tridiagonalization, the Cholesky factorization, and conjugate gradient algorithm (CGA); see Section 3.1 for a brief summary of these algorithms. First, we apply Theorem 2.2 to these algorithms and obtain accurate asymptotic formulae for the key quantities. For Lanczos, it is equivalent to the study of the asymptotics of the three-term recurrence coefficients of the (discrete) orthogonal polynomials. The results are recorded in Corollary 3.2. The Cholesky factorization of the Lanczos Jacobi matrix (cf. (3.5)) can also be analyzed similarly as in Corollary 3.4. This Cholesky factorization coincides with the well-known Golub–Kahan bidiagonalization procedure, which, as pointed out previously, has a full distributional characterization in the isotropic Gaussian case. But our results hold for non-Gaussian samples with non-trivial covariance. CGA is analyzed in Corollary 3.3. Based on the unperturbed asymptotics for μ, we establish the perturbed asymptotics for these algorithms and the results are reported in Theorem 3.6. Again, the leading errors can be fully expressed in terms of (1.4) and the associated theta functions.

As mentioned earlier, the fluctuations of the perturbed orthogonal polynomials and related quantities of the numerical algorithms depend on (1.4), which should be expected to have a problem-specific form. In Section 4, we consider a concrete case study, in the random matrix context, using a general spiked sample covariance matrix model. More specifically, ν is the VESD of the $N \times N$ sample covariance matrix whose deterministic equivalent μ can be characterized using the anisotropic local laws as discussed in Section 4.3. The methodology we propose here shows how Riemann–Hilbert problems can assist yet again, later in the analysis of a random matrix ensemble, once one has some knowledge of the local law. The main result is Theorem 4.3, which establishes a general mesoscopic-type central limit theorem (CLT) by analyzing a functional version of (1.4). We mention that the CLT is mesoscopic as its scaling also depends n. Informally, we prove that for $z \in \mathbb{R}$, when $n \ll N^{1/6}$

$$\frac{\sqrt{N/n^2}}{\mathsf{Z}(z;\mu)}(\pi_n(z;\mu) - \pi_n(z;\nu)) \xrightarrow[N\to\infty]{(d)} \mathcal{N}(0, \mathsf{d}(z)(\mathsf{V}_1 + \mathsf{V}_2)),$$

where $\mathsf{Z}(z;\mu)$ is a normalization constant that depends on z and μ, V_1 depends on μ and is independent of n, V_2 depends on both n and the fourth moments of the entries of the matrix, and $\xrightarrow[N\to\infty]{(d)}$ indicates convergence in law. Moreover, as long as $n \to \infty$, $\mathsf{V}_2 \to 0$ so that the CLT only depends on the first two moments. Finally, $\mathsf{d}(z)$ is a deterministic function depending on the application under consideration. For example, for the various aforementioned numerical algorithms, $\mathsf{d}(z)$ can be found explicitly is summarized in Corollary 4.4. Nevertheless, we mention that even though we work on the spiked sample covariance matrix model in the current paper, our methods can be easily applied to other random matrix models once the local laws are established.

We emphasize that our results of the case study generalize many existing results in numerical linear algebra and random matrix theory. First, we show that for a general class of spiked sample covariance matrices, if $n \ll N^{1/6}$ then the upper-left $n \times n$ subblock of L in (1.2) tends to the upper-left subblock of the Cholesky factorization of the three-term recurrence Jacobi matrix for the orthogonal polynomials with respect to the limiting VESD, with universal Gaussian fluctuations. We also establish that the dependence on the fourth moment diminishes as n increases, a phenomenon that was empirically observed in [64]. Second, we establish precise convergence statistics for

CGA when the matrix is a general spiked sample covariance matrix model. We allow n, which here is taken to be the number of iterations in CGA, to be divergent with N. In particular, we show that the residuals always have Gaussian fluctuations and become more universal (i.e., only depend on the first two moments) as more iterations are run. Comparable results have only been previously established for fixed n and trivial covariance case in [64] for the case of Wishart matrices.

Finally, we highlight an open question. In the current paper, the breakthrough allows n to increase with N in a moderate way, that is, $1 \leq n \leq N^{\alpha}, 0 \leq \alpha < 1/6$. It is interesting to consider the regime $1/6 \leq \alpha \leq 1$. Based on our numerical simulations, we conjecture that our results still hold for all $0 \leq \alpha < 1$. However, when $\alpha = 1$, our current results clearly fail to hold (see Figure 5) and we need to develop entirely new tools to handle this regime. We will pursue this direction in the future.

Conventions. For two sequences of real values $\{a_N\}$ and $\{b_N\}$, we write $a_N = \mathrm{O}(b_N)$ if $|a_N| \leq C|b_N|$ for some constant $C > 0$, and $a_N = \mathrm{o}(b_N)$ if $|a_N| \leq c_N|b_N|$ for some positive sequence $c_N \downarrow 0$. Moreover, we write $a_N \asymp b_N$ if $a_N = \mathrm{O}(b_N)$ and $b_N = \mathrm{O}(a_N)$. The notation $\langle \mathbf{b}, \mathbf{a} \rangle$ is used for the standard ℓ^2 inner product and $\|\mathbf{b}\|_2^2 = \langle \mathbf{b}, \mathbf{b} \rangle$. We use $\mathbf{f}_k$ to denote the kth standard Euclidean basis vector.

2 The Riemann–Hilbert Problem for Orthogonal Polynomials and Their Perturbations

Consider a probability measure μ without a singular continuous part. We suppose its absolute continuous density ρ is supported on a finite number of disjoint intervals $[\mathrm{a}_j, \mathrm{b}_j], 1 \leq j \leq g+1$. We also allow μ having a finite number of spikes, that is, point masses at $\mathrm{c}_i, 1 \leq i \leq p$, with masses w_j.

In [41], the authors found a characterization of orthogonal polynomials in terms of a matrix Riemann–Hilbert problem. We now review such a formulation. Define the Cauchy transforms of the monic polynomials

$$c_n(z;\mu) = \frac{1}{2\pi \mathrm{i}} \int_{\mathbb{R}} \frac{\pi_n(\lambda;\mu)}{\lambda - z} \mu(\mathrm{d}\lambda), \tag{2.1}$$

and the matrix-valued function

$$Y_n(z;\mu) = \begin{bmatrix} \pi_n(z;\mu) & c_n(z;\mu) \\ \gamma_{n-1}(\mu)\pi_{n-1}(z;\mu) & \gamma_{n-1}(\mu)c_{n-1}(z;\mu) \end{bmatrix}, \quad z \notin \mathrm{supp}(\mu), \tag{2.2}$$

where we used the notation

$$\gamma_n(\mu) = -2\pi \mathrm{i} \|\pi_n(\cdot;\mu)\|_{L^2(\mu)}^{-2}. \tag{2.3}$$

It then follows that (see [41] or [57])

$$Y_n^+(z;\mu) = Y_n^-(z;\mu) \begin{bmatrix} 1 & \rho(z) \\ 0 & 1 \end{bmatrix}, \quad Y_n^{\pm}(z;\mu) := \lim_{\epsilon \to 0^+} Y_n(z \pm \mathrm{i}\epsilon;\mu), \tag{2.4}$$

at all points $z \in \mathbb{R}$ where μ has a continuous density ρ. Additionally,

$$Y_n(z;\mu) \begin{bmatrix} z^{-n} & 0 \\ 0 & z^n \end{bmatrix} = I + \mathrm{O}(1/z), \quad z \to \infty. \tag{2.5}$$

Due to the discrete contributions to μ, this does not fully characterize Y_n. We compute

$$\begin{aligned} \mathrm{Res}_{z=\mathrm{c}_j} Y_n(z;\mu) &= \begin{bmatrix} 0 & \frac{1}{2\pi \mathrm{i}} w_j \pi_n(\mathrm{c}_j;\mu) \\ 0 & \frac{\gamma_{n-1}}{2\pi \mathrm{i}} w_j \pi_{n-1}(\mathrm{c}_j;\mu) \end{bmatrix} \\ &= \lim_{z \to \mathrm{c}_j} Y_n(z;\mu) \begin{bmatrix} 0 & \frac{w_j}{2\pi \mathrm{i}} \\ 0 & 0 \end{bmatrix}, \quad j = 1, 2, \ldots, p. \end{aligned} \tag{2.6}$$

Conditions (2.4), (2.5), and (2.6) constitute a Riemann–Hilbert problem for $Y_n(z;\mu)$ and $Y_n(z;\mu)$ is the unique solution of this problem if one requires continuous boundary values.

Remark 2.1. At points where μ has a density, but it fails to be continuous, one may have to impose additional conditions to uniquely characterize Y_n. The assumptions we impose on μ in the current work allow us to ignore such complications.

2.1 Perturbation theory for orthogonal polynomials

Let ν be a perturbed (and potentially random) version of μ. Suppose μ and ν are both measures supported on a finite number (i.e., $g+1$) of intervals with a finite number (i.e.,

p) of spikes for (potentially) different choices of $\mathrm{a}_j, \mathrm{b}_j, w_j, \mathrm{c}_j, h_j$, and g, p. Define

$$\tilde{Y}_n(z;\mu) = \begin{cases} Y_n(z;\mu)\begin{bmatrix} 1 & -c_0(z;\mu) \\ 0 & 1 \end{bmatrix} & z \text{ inside } \Gamma, \\ Y_n(z;\mu) & \text{otherwise,} \end{cases} \tag{2.7}$$

where Γ is a simple curve with counter-clockwise orientation that encloses the support of μ. Using (2.4), we then compute the jumps of $\tilde{Y}_n$ on $\cup_j(\mathrm{a}_j, \mathrm{b}_j)$:

$$\begin{aligned} \tilde{Y}_n^+(z;\mu) &= Y_n^+(z;\mu)\begin{bmatrix} 1 & -c_0^+(z;\mu) \\ 0 & 1 \end{bmatrix} = Y_n^-(z;\mu)\begin{bmatrix} 1 & \rho(z) \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & -c_0^+(z;\mu) \\ 0 & 1 \end{bmatrix} \\ &= Y_n^-(z;\mu)\begin{bmatrix} 1 & \rho(z) - c_0^+(z;\mu) \\ 0 & 1 \end{bmatrix} = \tilde{Y}_n^-(z;\mu)\begin{bmatrix} 1 & c_0^-(z;\mu) + \rho(z) - c_0^+(z;\mu) \\ 0 & 1 \end{bmatrix}. \end{aligned}$$

For $z \in \cup_j[\mathrm{a}_j, \mathrm{b}_j]$ the inversion formula holds [3], that is,

$$c_0^+(z;\mu) - c_0^-(z;\mu) = \rho(z),$$

and therefore $\tilde{Y}_n$ has a trivial jump on $\cup_j(\mathrm{a}_j, \mathrm{b}_j)$. Next, using (2.2) and residue theorem, we check the residues of $\tilde{Y}_n(z;\mu)$

$$\begin{aligned} \operatorname{Res}_{z=\mathrm{c}_j}\tilde{Y}_n(z;\mu) &= \operatorname{Res}_{z=\mathrm{c}_j} Y_n(z;\mu)\begin{bmatrix} 1 & -c_0(z;\mu) \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 0 & \operatorname{Res}_{z=\mathrm{c}_j}(-c_0(z;\mu)(Y_n(z;\mu))_{11} + (Y_n(z;\mu))_{12}) \\ 0 & \operatorname{Res}_{z=\mathrm{c}_j}(-c_0(z;\mu)(Y_n(z;\mu))_{21} + (Y_n(z;\mu))_{22}) \end{bmatrix} = 0. \end{aligned}$$

We conclude that $\tilde{Y}_n(z;\mu)$ must be analytic inside Γ and satisfies

$$\tilde{Y}_n^+(z;\mu) = \tilde{Y}_n^-(z;\mu)\begin{bmatrix} 1 & -c_0(z;\mu) \\ 0 & 1 \end{bmatrix}, \quad z \in \Gamma, \tag{2.8}$$

$$\tilde{Y}_n(z;\mu)\begin{bmatrix} z^{-n} & 0 \\ 0 & z^n \end{bmatrix} = I + \mathrm{O}(1/z), \quad z \to \infty. \tag{2.9}$$

As we will see in the next section, it is convenient to consider

$$\check{Y}_n(z;\mu) = \mathfrak{c}^{(n-p)\sigma_3}\tilde{Y}_n(z;\mu), \quad \mathfrak{c} \in \mathbb{C}\setminus\{0\}, \quad \sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \tag{2.10}$$

where $\mathfrak{c}$ is closely related to the capacity of $\cup_i[\mathrm{a}_i, \mathrm{b}_i]$ and formally defined in (A.10) after necessary notation is introduced. Note that the above modification does not affect the jump satisfied by $\check{Y}_n$, only its asymptotics.

To connect the two measures, μ and ν, we consider

$$X_n(z;\mu,\nu) = \check{Y}_n(z;\nu)\check{Y}_n(z;\mu)^{-1}, \tag{2.11}$$

where we note that $\det \check{Y}_n(z;\mu) \equiv 1$. Using (2.8) and (2.9), by an elementary calculation,

$$X_n^+(z;\mu,\nu) = X_n^-(z;\mu,\nu)J_n(z;\mu,\nu),\ z \in \Gamma;\ \text{and}\ \ X_n(z;\mu,\nu) = I + \mathrm{O}(1/z),\ z \to \infty,$$

where $J_n(z;\mu,\nu)$ is defined as

$$J_n(z;\mu,\nu) := \left[I + c_0(z,\mu-\nu)\check{Y}_n^-(z;\mu)\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\check{Y}_n^-(z;\mu)^{-1} \right].$$

Now, suppose that $\Gamma = \Gamma(N)$, $\nu = \nu(N)$ and $n = n(N)$ depend on a common asymptotic parameter N. The Riemann–Hilbert problem for X_n can be reformulated as a singular integral equation for a new unknown U_n defined on Γ using the representation

$$X_n(z;\mu,\nu) = I + \mathcal{C}_\Gamma U_n(z;\mu,\nu), \quad \mathcal{C}_\Gamma U(z) := \frac{1}{2\pi \mathrm{i}}\int_\Gamma \frac{U(z')}{z'-z}\mathrm{d}z'.$$

Proposition 2.1. For an integer N, suppose $\Gamma = \Gamma(N)$ is a piecewise smooth, simple, closed curve that encircles $\mathrm{supp}(\mu)\cup\mathrm{supp}(\nu)$ such that the operator norm of $\mathcal{C}_\Gamma^-$ on $L^2(\Gamma)$ is bounded by C_N. Suppose $n = n(N)$ and $\nu = \nu(N)$ are functions of N such that as $N \to \infty$, $C_N\|J_n - I\|_{L^\infty(\Gamma)} \to 0$. Then we have

$$X_n(z;\mu,\nu) = I + \frac{1}{2\pi\mathrm{i}}\int_\Gamma \frac{c_0(z';\mu-\nu)M_n(z';\mu)}{z'-z}\mathrm{d}z' + \mathrm{O}\left(C_N\frac{\|J_n - I\|^2_{L^\infty(\Gamma)}}{1+|z|}\right), \tag{2.12}$$

$$M_n(z;\mu) = \check{Y}_n^-(z;\mu)\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\check{Y}_n^-(z;\mu)^{-1}, \tag{2.13}$$

uniformly on subsets of $\mathbb{C}$ bounded uniformly away from Γ.

Proof. Define the boundary-value operator $\mathcal{C}_\Gamma^{\pm} U(z) = \lim_{z' \to z} \mathcal{C}_\Gamma U(z')$, where the limit is taken non-tangentially within the interior (+) or exterior (−) of Γ. Then U_n must satisfy

$$U_n - \mathcal{C}_\Gamma^- U_n (J_n - I) = J_n - I.$$

This is a near-identity operator equation for N sufficiently large and it can therefore be solved by a Neumann series. In particular,

$$\|U_n - (J_n - I)\|_{L^2(\Gamma)} = \mathrm{O}(C_N \|J_n - I\|^2_{L^\infty(\Gamma)}),$$

which implies the conclusion. ■

Remark 2.2. Proposition 2.1 establishes the perturbation for orthogonal polynomials generated by two close measures using the quantity (2.11). In particular, let

$$P(z;n) = X_n(z;\mu,\nu) - I. \tag{2.14}$$

Using (2.12) and the definition (2.11), we readily see that

$$\begin{aligned} \pi_n(z;\nu) &= \pi_n(z;\mu)(1 + P_{11}(z;n)) + \mathfrak{c}^{2(p-n)} \gamma_{n-1}(\mu) \pi_{n-1}(z;\mu) P_{12}(z;n), \\ c_n(z;\nu) &= c_n(z;\mu)(1 + P_{11}(z;n)) + \mathfrak{c}^{2(p-n)} \gamma_{n-1}(\mu) c_{n-1}(z;\mu) P_{12}(z;n), \end{aligned} \tag{2.15}$$

where P_{ij} is the (i,j) entry of P. If the two measures are close, the functions P_{ij} will decay so that, to leading order, $\pi_n(z;\nu)$ and $c_n(z;\nu)$ are given by $\pi_n(z;\mu)$ and $c_n(z;\mu)$, as expected. The above results may depend on the choice of the contour Γ. In the current paper, we will choose Γ to be the boundary of a rectangle and $\|\mathcal{C}_\Gamma^-\|_{L^2(\Gamma)}$ is bounded by an absolute constant [14].

2.2 Large n asymptotics of polynomials orthogonal with respect to measures supported on multiple intervals

Recall (2.2). In order to directly compare the orthogonal polynomial $\pi_n(x;\nu)$ to $\pi_n(x;\mu)$ one needs (1) an estimate on $M_n(z;\mu)$ in (2.13). Furthermore, supposing that $J_n - I \to 0$, one is left with

$$\check{Y}_n(z;\nu) = X_n(z;\mu,\nu)\check{Y}_n(z;\mu) = (I + \mathrm{o}(1))\check{Y}_n(z;\mu).$$

And so, one needs (2) some information about $\check{Y}_n(z;\mu)$ to make conclusions about $\check{Y}_n(z;\nu)$. One such way to accomplish (1) and (2) is to compute the large n asymptotics of $Y_n(z;\mu)$. The calculations rely on solving another Riemann–Hilbert problem, and this is accomplished in Appendix A. We summarize the results in Theorem 2.2 below. The result relies on the following regularity assumption.

Assumption 1. Consider a probability measure μ that satisfies the following assumptions.

(1) Square-root behavior with spikes: The measure μ is of the form (One can include inverse square-roots if needed, but this requires incorporating additional conditions into the Riemann–Hilbert problem to ensure unique solvability.)

$$\mu(\mathrm{d}\lambda) = \underbrace{\sum_{j=1}^{g+1} h_j(\lambda)\mathbf{1}_{[\mathrm{a}_j,\mathrm{b}_j]}(\lambda)\sqrt{(\mathrm{b}_j-\lambda)(\lambda-\mathrm{a}_j)}\,\mathrm{d}\lambda}_{\rho(\lambda)} + \sum_{j=1}^{p} w_j\delta_{\mathrm{c}_j}(\mathrm{d}\lambda), \tag{2.16}$$

for disjoint intervals $[\mathrm{a}_j,\mathrm{b}_j]$ and points c_j located away from these intervals.

(2) Uniformity (1): We allow μ to depend implicitly on a parameter N but require that g, p be non-negative, constant (for sufficiently large N) and require that the distance between any two points in the set $\{\mathrm{c}_j\}\cup\{\mathrm{a}_j\}\cup\{\mathrm{b}_j\}$ is bounded above and below.

(3) Analyticity: To each interval $[\mathrm{a}_j,\mathrm{b}_j]$, we associate a bounded open set Ω_j (independent of N) containing $[\mathrm{a}_j,\mathrm{b}_j]$ for all N such that h_j has an analytic continuation to Ω_j.

(4) Uniformity (2): We suppose there is an absolute constant $D\geq 1$ such that

$$\sup_{z\in\Omega_j}\max\{|h_j(z)|,|h_j(z)|^{-1}\}\leq D,$$

for every $1\leq j\leq g+1$.

(5) Uniformity (3): For every j, we assume that either $N^{-\sigma}/D\leq|w_j|\leq D$, $0\leq\sigma<\infty$, or $w_j=0$.

We point out that the limiting ESDs and VESDs for many commonly studied random matrix models satisfy Assumption 1. We refer the readers to Lemma 4.1 and the discussion below for more details on this. Now we state the results. Let D_j be a small region containing $[\mathrm{a}_j, \mathrm{b}_j]$, and let $\mathring{\Sigma}_j$ be a small ball that has c_j as its center. Then we define a function f as follows:

$$f(z) = \begin{cases} \pm 1/\check{\rho}_j(z) & z \in D_j \cap \{\pm \mathrm{Im}\, z > 0\}, \\ \frac{\tilde{w}_j}{z - \mathrm{c}_j} & z \in \mathring{\Sigma}_j, \\ 0 & \text{otherwise,} \end{cases}$$

where $\tilde{w}_j$ is defined in (A.6) and $\check{\rho}_j$ is defined in Section A.2.3 after necessary notation is introduced. Since D_j and $\mathring{\Sigma}_j$ can be chosen to be well separated according to Assumption 1, we will see in Section A.2.3 that their choices will not influence our results much. The function f here captures the fact that the asymptotics for orthogonal polynomials away from the support of μ is different from the asymptotics on or near the support.

Theorem 2.2. Suppose Assumption 1 holds for $\mu = \mu(N)$ for sufficiently large N. Let $Y_n(z;\mu)$ be as (2.2) and recall $\mathfrak{c}, \sigma_3$ in (2.10). Then for some constant $c > 0$

$$Y_n(z;\mu) = \mathfrak{c}^{(p-n)\sigma_3}\left(I + \mathrm{O}\left(\frac{\mathrm{e}^{-cn}}{1+|z|}\right)\right) K_n(z,\mu)\, \mathrm{e}^{\varphi_n(z)\sigma_3} \begin{bmatrix} 1 & 0 \\ f(z) & 1 \end{bmatrix} \left(\prod_{j=1}^{p}(z - \mathrm{c}_j)\right)^{\sigma_3}. \tag{2.17}$$

Here, we used the notation

$$K_n(z,\mu) = \mathrm{e}^{-\sigma_3 \mathsf{G}(\infty)} L_n(\infty)^{-1} L_n(z), \tag{2.18}$$

$$\varphi_n(z) = \mathsf{G}(z) + (n-p)\mathfrak{g}(z), \tag{2.19}$$

where $\mathsf{G}(z)$ is defined (A.11), $L_n(z)$ is defined in (A.5), and $\mathfrak{g}(z)$ is defined in Section A.2.2, after some necessary notation is introduced.

Proof. See Appendix A. ■

The function $\mathfrak{g}(z)$, as defined in Section A.2.2, is classically known as the exterior Green's function with pole at ∞, see [65], for example. It expresses the global distribution of the zeros of the orthogonal polynomials. The function $\mathsf{G}(z)$ is an instance of a so-called Szegoő function [55]. For the definition of L_n see (B.6).

For the reader's convenience, in Appendix B.1, we provide more detailed expressions for the entries of $Y_n(z;\mu)$. Theorem 2.2 has many important consequences. For example, it can be used to study the asymptotics of the three-term recurrence coefficients of the orthogonal polynomials (see Section 3.2.1), the residuals and errors of conjugate gradient algorithm (see Section 3.2.2) and the Cholesky factorization of the tridiagonalization (see Section 3.2.3). We will discuss these applications and provide explicit formulae in Section 3.2.

Equipped with the above theorem, we now proceed to accomplish the aforementioned goals (1) and (2) on some specifically chosen contour Γ. In sequel, unless otherwise specified, we will consistently use the following contour. For some small constant $\eta > 0$, let Γ_j be the rectangle that is a distance η from $[\mathsf{a}_j, \mathsf{b}_j]$, that is,

$$\Gamma_j = \Gamma_j(\eta) = \left([\mathsf{a}_j - \eta, \mathsf{b}_j + \eta] + \mathrm{i}\eta\right) \cup \left([\mathsf{a}_j - \eta, \mathsf{b}_j + \eta] - \mathrm{i}\eta\right) \cup \left(\mathsf{b}_j + \eta + \mathrm{i}[-\eta, \eta]\right) \cup \left(\mathsf{a}_j - \eta + \mathrm{i}[-\eta, \eta]\right). \tag{2.20}$$

The following lemma accomplishes (1) by providing an estimate on $M_n(z; \mu)$ in (2.13). For definiteness, we consider the matrix norm $\|A\|_{\max} = \max_{ij} |A_{ij}|$.

Lemma 2.3. Suppose Assumption 1 holds. We have that for $z \in \Gamma_j$ in (2.20), uniformly,

$$\|M_n(z; \mu)\|_{\max} \le C\eta^{-1}\, \mathrm{e}^{C' n \eta^{1/2}}, \tag{2.21}$$

for constants $C, C' > 0$.

Proof. We start by preparing some basic estimates. First, on Γ_j, according to Assumption 1, we have

$$C^{-1}\eta \le \prod_{j=1}^{g+1} |z - \mathsf{a}_j| \le C, \quad C^{-1}\eta \le \prod_{j=1}^{g+1} |z - \mathsf{b}_j| \le C,$$

for an absolute constant $C > 0$. Second, using (A.5) and (A.3) together with (A.12), we see from (2.18) that for $z \in \Gamma_j$, uniformly,

$$\|K_n(z; \mu)\|_{\max} \le C(|z - \mathsf{a}_j|^{-1/4} + |z - \mathsf{b}_j|^{-1/4}),$$
$$\|K_n(z; \mu)^{-1}\|_{\max} \le C(|z - \mathsf{a}_j|^{-1/4} + |z - \mathsf{b}_j|^{-1/4}),$$

for some absolute constant $C > 0$. Third, to estimate $\mathfrak{g}(z)$ in the upper-half plane, we first note that (Here $^+$ denotes the limit from within the upper-half plane.) $\operatorname{Re}\mathfrak{g}^+(z) = 0$ for

$z \in [\mathrm{a}_j, \mathrm{b}_j]$ for any j. According to the arguments of Section A.2.2, we find that there exists some $D > 0$ such that $|\Omega_g(z)| \leq D$ (recall (A.7)) on $\cup_j \Gamma_j$, which implies that for $z \in \Gamma_j$

$$\operatorname{Re} \mathfrak{g}(z) \leq D' \operatorname{dist}(z, [\mathrm{a}_j, \mathrm{b}_j])^{1/2} \leq 2^{1/4} D' \eta^{1/2},$$

for a new absolute constant $D' > 0$.

Next, we estimate $M_n(z; \mu)$. Inserting (2.17) into (2.13), we obtain

$$M_n(z;\mu) = \prod_{j=1}^{p} (z - \mathrm{c}_j)^2 \left(I + \mathrm{O}\left(e^{-cn}\right)\right) K_n(z;\mu) \begin{bmatrix} -f(z) & \mathrm{e}^{2\varphi_n(z)} \\ -f(z)^2\, \mathrm{e}^{-2\varphi_n(z)} & f(z) \end{bmatrix} K_n(z;\mu)^{-1} \left(I + \mathrm{O}\left(e^{-cn}\right)\right).$$

Using Lemma A.2, we estimate for $z \in \Gamma_j$

$$|\,\mathrm{e}^{2\varphi_n(z)}\,| \|K_n(z;\mu)\|_{\max} \|K_n(z;\mu)^{-1}\|_{\max} \leq C|z - \mathrm{b}_j|^{-1}|z - \mathrm{a}_j|^{-1},$$

$$|f(z)| \|K_n(z;\mu)\|_{\max} \|K_n(z;\mu)^{-1}\|_{\max} \leq C|z - \mathrm{b}_j|^{-1}|z - \mathrm{a}_j|^{-1},$$

$$|f(z)|^2 |\,\mathrm{e}^{-2\varphi_n(z)}\,| \|K_n(z;\mu)\|_{\max} \|K_n(z;\mu)^{-1}\|_{\max} \leq C|z - \mathrm{b}_j|^{-1}|z - \mathrm{a}_j|^{-1},$$

for a new constant $C > 0$. The lemma follows. ■

Armed with Lemma 2.3, we are ready to state a more detailed asymptotic result on the perturbation of orthogonal polynomials when Assumption 1 holds.

Theorem 2.4. Let N be a positive integer and suppose $\mu = \mu(N)$ satisfies Assumption 1 for sufficiently large N. Suppose further that a measure $\nu = \nu(N)$ is such that

$$\nu - \sum_{j=1}^{p} w_j \delta_{\mathrm{c}_j},$$

has its support inside $\Gamma = \Gamma(\eta) = \bigcup_j \Gamma_j(\eta)$, as defined in (2.20), and $\|c_0(\cdot, \mu - \nu)\|_{L^\infty(\Gamma)} \leq E(N, \eta)$. If $n \leq C\eta^{-1/2}$, $C > 0$, and $\eta = \eta(N)$ is such that $E(N,\eta)\eta^{-1/2} \to 0$ as $N \to \infty$, then Proposition 2.1 holds. In particular, we have

$$X_n(z;\mu,\nu) = I + \frac{1}{2\pi \mathrm{i}} \int_\Gamma \frac{c_0(z';\mu - \nu) M_n(z';\mu)}{z' - z} \mathrm{d}z' + \mathrm{O}\left(\frac{E(N,\eta)^2 \eta^{-1}}{1 + |z|}\right),$$

uniformly for z in sets bounded away from Γ.

Proof. The proof follows directly from Theorem 2.2, Lemma 2.3, and Proposition 2.1. ■

Remark 2.3. Theorem 2.4 makes precise the fact that in order to let $X_n(z;\mu,\nu)$ be close to I, we will need $c_0(z;\mu-\nu)$ to be small. The sense in which this occurs depends on each specific problem and the related application. In applications of random matrix theory, for most of the commonly encountered models, when μ is the limiting ESD or VESD and ν is the ESD or VESD, one typically has ($X_n = \mathrm{O}_{\mathbb{P}}(g(n))$ as $n \to \infty$ if $|c_n X_n/g(n)| \to 0$ in probability for any sequence $c_n \to 0$.)

$$|c_0(z;\mu-\nu)| = \mathrm{O}_{\mathbb{P}}\left(\frac{1}{N\eta}\right), \quad \text{or } |c_0(z;\mu-\nu)| = \mathrm{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{N\eta}}\right),$$

on the entirety of $\cup_j \Sigma_j$ and this will dictate what η, or equivalently n, can be. Consequently, we choose

$$n = \mathrm{O}(\eta^{-1/2}), \text{ where } n \ll N^{1/4} \text{ for ESD and } n \ll N^{1/6} \text{ for VESD,} \tag{2.22}$$

is required to be able to apply Theorem 2.4. We also point out that if μ has spikes, then ν will have spikes near the spikes of μ. Instead of directly considering $\mu-\nu$ we apply Theorem 2.4 to $\tilde{\mu}-\nu$ where the limiting spikes of μ are replaced with the nearby random spikes of ν. Despite the fact that $\tilde{\mu}$ is then random, it satisfies Assumption 1 with high probability and the asymptotics of the associated orthogonal polynomials follow the same form, see Remark 2.4 below.

Remark 2.4. Combining Theorems 2.2 and 2.4, we can provide a more detailed perturbation formulae for the orthogonal polynomials compared to (2.15). In particular, inserting (2.17) (or equivalently the expressions in Appendix B.1) into (2.15), we obtain that for z bounded away from Γ,

$$\begin{aligned}
\pi_n(z;\nu) &= \mathfrak{c}^{(p-n)}\, \mathrm{e}^{(n-p)\mathfrak{g}(z)+\mathsf{G}(z)-\mathsf{G}(\infty)} \\
&\quad \times \left[\prod_{j=1}^{p}(z-\mathsf{c}_j)\right]\Big[(1+P_{11}(z;n))E_{11}(z;n)+P_{12}(z;n)\,\mathrm{e}^{2\mathsf{G}(\infty)}\,E_{21}(z;n)\Big],\\
c_n(z;\nu) &= \mathfrak{c}^{(p-n)}\, \mathrm{e}^{-(n-p)\mathfrak{g}(z)-\mathsf{G}(z)-\mathsf{G}(\infty)} \\
&\quad \times \left[\prod_{j=1}^{p}(z-\mathsf{c}_j)^{-1}\right]\Big[(1+P_{11}(z;n))E_{12}(z;n)+P_{12}(z;n)\,\mathrm{e}^{2\mathsf{G}(\infty)}\,E_{22}(z;n)\Big],
\end{aligned} \tag{2.23}$$

where $E_{ij}(z;n), 1 \leq i,j \leq 2$, defined in Appendix B.1 only depend on μ. Compared to (2.15), the above expressions give much more information as they give explicitly how the exponential prefactors are arranged.

Remark 2.5. As can be seen from the above discussion, if ν is random then the main random quantity to be understood is the entries of $P(z;n)$ as defined in (2.14). Consequently, in order to understand the second-order fluctuation of the concerned quantities, it suffices to derive a CLT for $P(z;n)$. The main task is to understand the asymptotics of $c_0(z';\mu-\nu)$ on the contour Γ. This is usually problem-specific and depends on the measures μ and ν. Considering applications in random matrix theory where μ is the limiting distribution and ν is the empirical distribution, the distribution of $c_0(z';\mu-\nu)$, of course, depends on the underlying random matrix model. In Section 4, we consider the spiked sample covariance matrix model and establish a general CLT, which can be used to understand the distribution of the related quantities.

3 Algorithmic Applications: Asymptotic Formulae for Numerical Algorithms

In this section, we apply the results of Section 2 to study several important numerical algorithms.

3.1 A high level discussion of matrix factorizations and algorithms

We briefly discuss background for the numerical algorithms under consideration.

3.1.1 *Lanczos tridiagonalization*

We first introduce the Householder tridiagonalizaton procedure. It is the process by which a real symmetric or complex Hermitian matrix W is transformed to a real symmetric tridiagonal matrix using Householder reflectors. Householder reflectors can be written in the form

$$U_k = \begin{bmatrix} I_k & 0 \\ 0 & I_{N-k} - 2\mathbf{u}\mathbf{u}^* \end{bmatrix},$$

where I_k is the $k \times k$ identity matrix and $\mathbf{u} \in \mathbb{C}^{(N-k)\times(N-k)}$ is a unit vector. By selecting $\mathbf{u}$ correctly for each k

$$U_N U_{N-1} \cdots U_1 W U_1^* U_2^* \cdots U_N^*,$$

is a real symmetric tridiagonal matrix. See [72], for example.

The Lanczos tridiagonalization algorithm applied to a real symmetric or complex Hermitian matrix W and vector $\mathbf{b}$ accomplishes the same goal as the Householder tridiagonalization algorithm with some added flexibility. Run to completion, in exact arithmetic, the Lanczos algorithm performs Gram–Schmidt on the vectors $\{\mathbf{b}, W\mathbf{b}, \ldots, W^{N-1}\mathbf{b}\}$ constructing an orthogonal or unitary matrix

$$Q = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \cdots & \mathbf{q}_N \end{bmatrix}, \tag{3.1}$$

and necessarily $T = Q^* W Q$ is a tridiagonal matrix. Note that $\mathbf{q}_1 = \mathbf{b}/\|\mathbf{b}\|_2$. It is well-known [72] the entries in the Lanczos matrix T coincides with the three-term recurrence coefficients for the discrete orthogonal polynomials with respect to the VESD generated by $\mathbf{b}$ and W (cf. (3.5)).

3.1.2 *Cholesky factorization*

The Cholesky factorization of a positive definite matrix W is a factorization $W = LL^*$, where L is lower-triangular with positive diagonal entries. When applied to a tridiagonal matrix T, L is lower-bidiagonal and has non-negative entries if T has non-negative entries. The Cholesky factorization is a special case of Gaussian elimination.

3.1.3 *The conjugate gradient algorithm*

The conjugate gradient algorithm (CGA) is an iterative method to solve the linear system $W\mathbf{x} = \mathbf{b}$. The method begins with an initial guess $\mathbf{x}_0$ and in the current work we always take $\mathbf{x}_0 = \mathbf{0}$. The algorithm is mathematically described by the solution of a sequence of minimization problems:

$$\mathbf{x}_k = \mathrm{argmin}_{\mathbf{y}\in\mathcal{K}_k} \|\mathbf{y} - \mathbf{x}\|_W, \quad \mathcal{K}_k = \mathrm{span}\{\mathbf{b}, W\mathbf{b}, \ldots, W^{k-1}\mathbf{b}\}, \quad \|\mathbf{y}\|_W^2 = \langle \mathbf{y}, W\mathbf{y}\rangle. \tag{3.2}$$

While one has the expression,

$$\mathbf{x}_k = Q_k (Q_k^* W Q_k)^{-1} \mathbf{f}_1,$$

it is quite remarkable that an extremely efficient iteration process is possible [46]. Here $Q_k := [\mathbf{q}_1, \cdots, \mathbf{q}_k]$ as in (3.1). It is also of intrinsic mathematical interest that this process makes sense for bounded positive-definite operators on a Hilbert space.

3.2 Unperturbed asymptotics: applications of Theorem 2.2

In this subsection, we consider several important consequences of Theorem 2.2 when applied to the numerical algorithms in Section 3.1. As we will see later, a common feature is that the analysis of these algorithms boil down to the analysis of some functionals of the orthogonal polynomials and Cauchy transforms evaluated at either $z = 0$ or $z = \infty$. The main theorem is now stated and its consequences follow.

Based on $\{\pi_n(\lambda;\mu)\}$ in (1.3), the orthonormal polynomials $p_n(\lambda;\mu)$, $n = 0, 1, 2, \ldots$, are the defined by

$$p_n(\lambda;\mu) = \frac{\pi_n(\lambda;\mu)}{\|\pi_n(\cdot;\mu)\|_{L^2(\mu)}}, \quad \|\pi_n(\cdot;\mu)\|^2_{L^2(\mu)} = \int_{\mathbb{R}} \pi_n(\lambda;\mu)^2 \mu(\mathrm{d}\lambda).$$

We write $p_n(z;\mu) = \ell_n z^n + s_n z^{n-1} + \cdots = \ell_n \pi_n(z;\mu)$ where $\ell_n = \ell_n(\mu)$ satisfies

$$\ell_n^{-2} = \int_{\mathbb{R}} \pi_n(z;\mu)^2 \mu(\mathrm{d}z) = \int_{\mathbb{R}} \pi_n(z;\mu) z^n \mu(\mathrm{d}z). \tag{3.3}$$

Theorem 3.1. Suppose Assumption 1 holds for $\mu = \mu(N)$ for sufficiently large N and $n \to \infty$ as $N \to \infty$. Then for some $c > 0$ we have the following.

(1) If $z = 0$ is bounded away from $\left(\cup_j \Omega_j\right) \cup \left(\cup_j \mathrm{c}_j\right)$ then (This result can be stated appropriately for any z but for simplicity we just take $z = 0$ because that is all that is needed in the sequel.)

$$Y_n(0;\mu)_{11} = \mathfrak{c}^{(p-n)}\, \mathrm{e}^{-\mathsf{G}(\infty)}\, \mathrm{e}^{\mathsf{G}(0)}\, \mathrm{e}^{(n-p)\mathfrak{g}(0)} \left[\prod_{j=1}^{p} (-\mathrm{c}_j)\right] E_{11}(0;n),$$

$$Y_n(0;\mu)_{12} = \mathfrak{c}^{(p-n)}\, \mathrm{e}^{-\mathsf{G}(\infty)}\, \mathrm{e}^{-\mathsf{G}(0)}\, \mathrm{e}^{-(n-p)\mathfrak{g}(0)} \left[\prod_{j=1}^{p} (-\mathrm{c}_j)^{-1}\right] E_{12}(0;n),$$

where

$$E_{11}(0;n) = \frac{1}{2}\left(\prod_{j=1}^{g+1}\left(\frac{\mathrm{b}_j}{\mathrm{a}_j}\right)^{1/4} + \prod_{j=1}^{g+1}\left(\frac{\mathrm{a}_j}{\mathrm{b}_j}\right)^{1/4}\right) \frac{\Theta_1(0;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})}{\Theta_1(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})} + \mathrm{O}(\mathrm{e}^{-cn}),$$

$$E_{12}(0;n) = \frac{1}{2\mathrm{i}}\left(\prod_{j=1}^{g+1}\left(\frac{\mathrm{b}_j}{\mathrm{a}_j}\right)^{1/4} - \prod_{j=1}^{g+1}\left(\frac{\mathrm{a}_j}{\mathrm{b}_j}\right)^{1/4}\right) \frac{\Theta_2(0;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})}{\Theta_1(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})} + \mathrm{O}(\mathrm{e}^{-cn}).$$

(2) And

$$\ell_n^{-2}(\mu) = -2\pi \mathrm{i} \lim_{z\to\infty} z^{n+1} Y_n(z;\mu)_{12}$$

$$= e^{-2\mathsf{G}(\infty)} \mathfrak{c}^{2(p-n)} \frac{\pi}{2} \sum_{j=1}^{g+1} (\mathsf{b}_j - \mathsf{a}_j) \frac{\Theta_2(\infty; \mathbf{d}_2; (n-p)\boldsymbol{\Delta} + \boldsymbol{\zeta})}{\Theta_1(\infty; \mathbf{d}_2; (n-p)\boldsymbol{\Delta} + \boldsymbol{\zeta})} + \mathrm{O}(\mathrm{e}^{-cn}),$$

$$\frac{s_n(\mu)}{\ell_n(\mu)} = \lim_{z\to\infty} z\left(z^{-n} Y_n(z;\mu)_{11} - 1\right)$$

$$= \frac{m_{g+1}}{2\pi \mathrm{i}} - \frac{m_g}{2\pi \mathrm{i}} \sum_{j=1}^{g+1} (\mathsf{a}_j + \mathsf{b}_j) + (n-p)\mathfrak{g}_1 - \sum_{j=1}^{p} \mathsf{c}_j + \frac{\Theta_1^{(1)}(\mathbf{d}_2; (n-p)\boldsymbol{\Delta} + \boldsymbol{\zeta})}{\Theta_1(\infty; \mathbf{d}_2; (n-p)\boldsymbol{\Delta} + \boldsymbol{\zeta})}$$

$$+ \mathrm{O}(\mathrm{e}^{-cn}).$$

Here $\mathfrak{c}$ is defined in (2.10) and $\mathfrak{g}_1$ is the coefficient of the $\mathrm{O}(1/z)$ term in the expansion of $\mathfrak{g}(z)$ at ∞. The other quantities will be made explicit in the proof after some necessary notation is introduced. In particular, $\Theta = (\Theta_1, \Theta_2)$ is a vector-valued function defined in (A.2) using the Riemann theta function (cf. (A.1)), $\mathbf{d}_2$ is defined in (A.4), $\boldsymbol{\Delta}$ is defined in (A.9), the entries of $\boldsymbol{\zeta}$ are defined via (A.13), and $\Theta^{(1)}$ is defined in (B.8).

Proof. See Appendix B.1. ∎

3.2.1 *Asymptotics of the three-term recurrence coefficients*

The three-term recurrence coefficients $a_n(\mu), b_n(\mu)$, $n \geq 0$, for $(p_n(x;\mu))_{n\geq 0}$ satisfy

$$a_n(\mu)p_n(x;\mu) + b_n(\mu)p_{n+1}(x;\mu) + b_{n-1}(\mu)p_{n-1}(x;\mu) = xp_n(x;\mu), \quad n \geq 0, \tag{3.4}$$

are often organized into a Jacobi matrix:

$$\mathcal{J}(\mu) = \begin{bmatrix} a_0 & b_0 & & & \\ b_0 & a_1 & b_1 & & \\ & b_1 & a_2 & b_2 & \\ & & b_2 & a_3 & \ddots \\ & & & \ddots & \ddots \end{bmatrix}, a_n = a_n(\mu), b_n = b_n(\mu). \tag{3.5}$$

We let $\mathcal{J}_n(\mu)$ denote the upper-left $n \times n$ subblock of $\mathcal{J}(\mu)$. The following theorem establishes the asymptotics of these coefficients.

Corollary 3.2. Suppose Assumption 1 holds for $\mu = \mu(N)$ for sufficiently large N. Then in the notation of Theorem 3.1 we have that

$$b_n(\mu)^2 = \frac{1}{\mathfrak{c}^2} \frac{\frac{\Theta_2(\infty;\mathbf{d}_2;(n+1)\boldsymbol{\Delta}+\boldsymbol{\zeta})}{\Theta_1(\infty;\mathbf{d}_2;(n+1)\boldsymbol{\Delta}+\boldsymbol{\zeta})} + \mathrm{O}(\mathrm{e}^{-cn})}{\frac{\Theta_2(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})}{\Theta_1(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})} + \mathrm{O}(\mathrm{e}^{-cn})},$$

$$a_n(\mu) = \frac{\Theta_1^{(1)}(\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})}{\Theta_1(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})} - \frac{\Theta_1^{(1)}(\mathbf{d}_2;(n+1)\boldsymbol{\Delta}+\boldsymbol{\zeta})}{\Theta_1(\infty;\mathbf{d}_2;(n+1)\boldsymbol{\Delta}+\boldsymbol{\zeta})} + \mathfrak{g}_1 + \mathrm{O}(\mathrm{e}^{-cn}).$$

Proof. See Appendix B.2. ∎

Remark 3.1. Two remarks are in order. First, Corollary 3.2 shows that the recurrence coefficients can be well approximated by some quantities involving the Riemann theta function (cf. (A.1) and (A.2) when $g > 0$. This, in turn, yields the approximate quasi-periodicity of $\{a_n\}$ and $\{b_n\}$. Second, we provide a single interval example to illustrate how different quantities in the above theorem can be calculated. In the general setting, these quantities can be calculated numerically, as will be discussed in Section 5.1. Consider that $g = 0$ and $p = 0$ in (2.16). When $\mathtt{b}_1 = 1$ and $\mathtt{a}_1 = -1$, one can check from (A.2) that $\Theta_1 = \Theta_2 = 1$ and $\mathfrak{g}_1 = 0$. Following [65], $\mathfrak{c}^{-2} = \frac{1}{4}$ so that

$$a_n = \mathrm{O}(\mathrm{e}^{-cn}),\ b_n = \frac{1}{2} + \mathrm{O}(\mathrm{e}^{-cn}),$$

which recovers the result of [55]. For general $\mathtt{a}_1$ and $\mathtt{b}_1$,

$$a_n = \frac{\mathtt{b}_1 + \mathtt{a}_1}{2} + \mathrm{O}(\mathrm{e}^{-cn}),\ b_n = \frac{\mathtt{b}_1 - \mathtt{a}_1}{4} + \mathrm{O}(\mathrm{e}^{-cn}),$$

which matches the result of [55] (see also [33, Theorem 5.2]).

3.2.2 *Asymptotics of CGA in infinite dimensions*

With the help of Corollary 3.2, we proceed to understand the performance of CGA (cf. (3.2)) to solve $W\mathbf{x} = \mathbf{b}$ with $\mathbf{x}_0 = \mathbf{0}$, producing iterates $\mathbf{x}_n$, $n = 1, 2, \ldots,$ and $\langle \mathbf{b}, (W - z)^{-1}\mathbf{b}\rangle = 2\pi \mathrm{i} c_0(z;\mu)$ for a measure μ. The residual and error vectors are defined as

$$\mathbf{r}_n = \mathbf{b} - W\mathbf{x}_n,\ \mathbf{e}_n = \mathbf{x} - \mathbf{x}_n.$$

Then we have the following formulae, where we note that for the assumptions of the theorem to hold, W must be an infinite-dimensional operator.

Corollary 3.3. Suppose Assumption 1 holds for $\mu = \mu(N)$ for sufficiently large N and $c_0(z;\mu) = 2\pi \mathrm{i}\langle \mathbf{b}, (W-z)^{-1}\mathbf{b}\rangle$. Then

$$\|\mathbf{e}_n\|_W^2 = 2\pi \mathrm{i}\, \mathrm{e}^{-2\mathrm{G}(0)}\, \mathrm{e}^{-2(n-p)\mathfrak{g}(0)} \left[\prod_{j=1}^{p} \mathrm{c}_j^{-2}\right] \frac{E_{12}(0;n)}{E_{11}(0;n)},$$

and

$$\|\mathbf{r}_n\|_2^2 = \frac{\frac{\pi}{2}\sum_{j=1}^{g+1}(\mathrm{b}_j - \mathrm{a}_j)\frac{\Theta_2(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})}{\Theta_1(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})} + \mathrm{O}(\mathrm{e}^{-cn})}{\mathrm{e}^{2(n-p)\mathfrak{g}(0)+2\mathrm{G}(0)}\left[\prod_{j=1}^{p}\mathrm{c}_j^2\right]E_{11}(0;n)^2}.$$

Here we recall the definitions of $G, \mathfrak{g}$ in (2.18) and (2.19), Θ in Theorem 3.1, and E_{11}, E_{12} are defined in Appendix B.1 after some necessary notation is introduced.

Proof. See Appendix B.2. ∎

Remark 3.2. As in Remark 3.1, the parameters of the above formulae can be calculated numerically as in Section 5.1. In the single interval case, together with (B.6) and (B.7), it is remarkable to see that

$$\frac{\|\mathbf{r}_n\|_2^2}{\|\mathbf{r}_{n-1}\|_2^2} = \mathrm{e}^{-2\mathfrak{g}(0)} + \mathrm{O}(\mathrm{e}^{-cn}), \quad \frac{\|\mathbf{e}_n\|_W^2}{\|\mathbf{e}_{n-1}\|_W^2} = \mathrm{e}^{-2\mathfrak{g}(0)} + \mathrm{O}(\mathrm{e}^{-cn}).$$

This implies that the ratios of the errors and residuals stay constant and are independent of the spikes. In fact, following the calculations in Section A.2.2, when $\mathrm{a}_1 > 0$ and $g = 0$ it is easy to see that $\mathrm{e}^{-\mathfrak{g}(0)} = (\sqrt{\mathrm{b}_1} - \sqrt{\mathrm{a}_1})/(\sqrt{\mathrm{b}_1} + \sqrt{\mathrm{a}_1})$, which matches [33, Theorem 3.3]. And in comparing with [29, 64] using the support $[(1-\sqrt{d})^2, (1+\sqrt{d})^2]$ of the Marchenko–Pastur distribution with parameter d, $0 < d \leq 1$, one obtains, for example,

$$\frac{\|\mathbf{r}_n\|_2^2}{\|\mathbf{r}_{n-1}\|_2^2} = d + \mathrm{O}(\mathrm{e}^{-cn}).$$

3.2.3 *Asymptotics of the Cholesky factorization*

It is well known that in the case where $\operatorname{supp}(\mu) \subset (0,\infty)$, the matrix $\mathcal{J}(\mu)$ in (3.5) has a Cholesky factorization

$$\mathcal{J}(\mu) = \mathcal{L}(\mu)\mathcal{L}(\mu)^*, \quad \mathcal{L}(\mu) = \begin{bmatrix} \alpha_0 & & & & \\ \beta_0 & \alpha_1 & & & \\ & \beta_1 & \alpha_2 & & \\ & & \beta_2 & \alpha_3 & \\ & & & \ddots & \ddots \end{bmatrix}, \quad \alpha_j = \alpha_j(\mu),\ \beta_j = \beta_j(\mu). \tag{3.6}$$

Let $\mathcal{L}_n(\mu)$ be the upper-left $n \times n$ subblock of $\mathcal{L}_n(\mu)$ and it is important that

$$\mathcal{J}_n(\mu) = \mathcal{L}_n(\mu)\mathcal{L}_n(\mu)^*.$$

The following holds.

Corollary 3.4. Suppose the assumptions of Theorem 2.2 hold, then we have that

$$\alpha_n(\mu)^2 = -\mathfrak{c}^{-1}\, \mathrm{e}^{\mathfrak{g}(0)} \frac{E_{11}(0; n+1)}{E_{11}(0; n)},$$
$$\beta_n(\mu)^2 = -\frac{\mathfrak{c} b_n(\mu)^2}{\mathrm{e}^{\mathfrak{g}(0)}} \frac{E_{11}(0; n)}{E_{11}(0; n+1)},$$

where the expansion of $b_n(\mu)$ can be found in Corollary 3.2.

Proof. See Appendix B.2. ■

Remark 3.3. First, as in Remark 3.2, in the single interval case $g = 0$, we can provide a more explicit formula. In this context, we have that

$$\alpha_n = \frac{\sqrt{\mathrm{a}_1} + \sqrt{\mathrm{b}_1}}{2} + \mathrm{O}(\mathrm{e}^{-cn}),\ b_n = \frac{\sqrt{\mathrm{b}_1} - \sqrt{\mathrm{a}_1}}{2} + \mathrm{O}(\mathrm{e}^{-cn}).$$

Second, according to [64, Section 6], we can also write

$$\frac{\|\mathbf{r}_n\|_2}{\|\mathbf{r}_{n-1}\|_2} = \frac{\beta_{n-1}}{\alpha_{n-1}}.$$

Combining the above two formulae will recover the arguments in Remark 3.2.

3.3 Perturbed formulae and perturbed asymptotics: Applications of Theorem 2.4

In this subsection, we consider several important consequences of Theorem 2.4 when applied to the aforementioned numerical algorithms. In what follows, we use ν as a perturbation of the measure μ and suppose that they satisfy the assumptions of Theorem 2.4. We first state how all the quantities that are analyzed in Theorem 3.1 are perturbed.

Theorem 3.5. For measures μ, ν satisfying the hypotheses of Proposition 2.1

$$Y_n(0;\nu)_{11} = Y_n(0;\mu)_{11}(1 + P_{11}(0;n)) + Y_n(0;\mu)_{21}P_{12}(0;n),$$

$$Y_n(0;\nu)_{12} = Y_n(0;\mu)_{12}(1 + P_{11}(0;n)) - 2\pi \mathrm{i}\frac{\mathfrak{c}^{2(p-n)}}{\ell_{n-1}^2(\mu)}Y_n(z;\mu)_{22}P_{12}(0;n),$$

$$\ell_n^{-2}(\nu) = \ell_n^{-2}(\mu) - 2\pi \mathrm{i}\mathfrak{c}^{2(p-n)}P_{12}^{(1)}(n),$$

$$\frac{\ell_n(\nu)}{s_n(\nu)} = \frac{\ell_n(\mu)}{s_n(\mu)} + P_{11}^{(1)}(n),$$

where the matrix $P(z;n) = P(z;n,\mu,\nu)$ is defined in (2.14) and $P^{(1)}(n) = P^{(1)}(n;\mu,\nu)$ is defined by

$$P^{(1)}(n) = \lim_{z\to\infty} zP(z;n). \tag{3.7}$$

Proof. This is a direct calculation first using

$$Y_n(z;\nu) = \mathfrak{c}^{(n-p)\sigma_3}(I + P(z;n))\mathfrak{c}^{(p-n)\sigma_3}Y_n(z;\mu),$$

and expanding

$$Y_n(z;\nu)z^{-n\sigma_3} = \mathfrak{c}^{(n-p)\sigma_3}(I + P(z;n))\mathfrak{c}^{(p-n)\sigma_3}Y_n(z;\mu)z^{-n\sigma_3},$$

in a series at infinity. ■

Since these are exact formulae, one can easily add the asymptotics of Theorem 3.1 (adding in the formulae (B.4) and (B.5)) to create perturbed versions of Corollaries 3.2, 3.3, and 3.4. We summarize this in the following theorem.

Theorem 3.6. Suppose the assumptions of Theorem 2.4 hold.

(1) For the three-term recurrence coefficients, corresponding to Corollary 3.2, we have

$$b_n(\nu)^2 = \frac{1}{\mathfrak{c}^2}\frac{\frac{\pi}{2}\sum_{j=1}^{g+1}(\mathrm{b}_j-\mathrm{a}_j)\frac{\Theta_2(\infty;\mathbf{d}_2;(n+1)\boldsymbol{\Delta}+\boldsymbol{\zeta})}{\Theta_1(\infty;\mathbf{d}_2;(n+1)\boldsymbol{\Delta}+\boldsymbol{\zeta})}+P^{(1)}_{12}(n+1)\,\mathrm{e}^{2G(\infty)}+\mathrm{O}(\mathrm{e}^{-cn})}{\frac{\pi}{2}\sum_{j=1}^{g+1}(\mathrm{b}_j-\mathrm{a}_j)\frac{\Theta_2(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})}{\Theta_1(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})}+P^{(1)}_{12}(n)\,\mathrm{e}^{2G(\infty)}+\mathrm{O}(\mathrm{e}^{-cn})},$$

and

$$a_n(\nu) = \frac{\Theta_1^{(1)}(\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})}{\Theta_1(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})} - \frac{\Theta_1^{(1)}(\mathbf{d}_2;(n+1)\boldsymbol{\Delta}+\boldsymbol{\zeta})}{\Theta_1(\infty;\mathbf{d}_2;(n+1)\boldsymbol{\Delta}+\boldsymbol{\zeta})}$$
$$+\,\mathfrak{g}_1 + P^{(1)}_{11}(n) - P^{(1)}_{11}(n+1) + \mathrm{O}(\mathrm{e}^{-cn}),$$

where the matrix $P^{(1)}$ is defined in (3.7).

(2) For CGA, corresponding to Corollary 3.3, we have

$$\|\mathbf{e}_n\|_W^2 = 2\pi\mathrm{i}\,\mathrm{e}^{-2(n-p)\mathfrak{g}(0)-2\mathrm{G}(0)}\left[\prod_{j=1}^{p}\mathrm{c}_j^{-2}\right]$$
$$\frac{(1+P_{11}(0;n))E_{12}(0;n)+P_{12}(0;n)\,\mathrm{e}^{2\mathrm{G}(\infty)}\,E_{22}(0;n)}{(1+P_{11}(0;n))E_{11}(0;n)+P_{12}(0;n)\,\mathrm{e}^{2G(\infty)}\,E_{21}(0;n)}, \tag{3.8}$$

$$\|\mathbf{r}_n\|_2^2$$
$$= \frac{\frac{\pi}{2}\sum_{j=1}^{g+1}(\mathrm{b}_j-\mathrm{a}_j)\frac{\Theta_2(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})}{\Theta_1(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})}+\frac{2\pi}{\mathrm{i}}P^{(1)}_{12}(n)\,\mathrm{e}^{2\mathrm{G}(\infty)}+\mathrm{O}(\mathrm{e}^{-cn})}{\mathrm{e}^{2(n-p)\mathfrak{g}(0)+2\mathrm{G}(0)}\left[\prod_{j=1}^{p}\mathrm{c}_j^{2}\right]\left[(1+P_{11}(0;n))E_{11}(0;n)+P_{12}(0;n)\,\mathrm{e}^{2\mathrm{G}(\infty)}\,E_{21}(0;n)\right]^2}. \tag{3.9}$$

(3) For the Cholesky factorization, corresponding to Corollary 3.4, we have

$$\alpha_n(\nu)^2$$
$$= -\mathfrak{c}^{-1}\,\mathrm{e}^{\mathfrak{g}(0)}\,\frac{(1+P_{11}(0;n+1))E_{11}(0;n+1)+P_{12}(0;n+1)\,\mathrm{e}^{2G(\infty)}\,E_{21}(0;n+1)}{(1+P_{11}(0;n))E_{11}(0;n)+P_{12}(0;n)\,\mathrm{e}^{2G(\infty)}\,E_{21}(0;n)},$$

$$\beta_n(\nu)^2$$
$$= -\mathfrak{c}\,\mathrm{e}^{-\mathfrak{g}(0)}\,b_n(\nu)^2\,\frac{(1+P_{11}(0;n))E_{11}(0;n)+P_{12}(0;n)\,\mathrm{e}^{2G(\infty)}\,E_{21}(0;n)}{(1+P_{11}(0;n+1))E_{11}(0;n+1)+P_{12}(0;n+1)\,\mathrm{e}^{2G(\infty)}\,E_{21}(0;n+1)}.$$

We do not present the formulae for $E_{12}(0;n)$ and $E_{22}(0;n)$ explicitly, but these can be found in Section B.1.

Remark 3.4. This theorem is particularly important because our asymptotic formulae in the previous section only hold when μ satisfies Assumption 1, which corresponds to running CGA on an infinite-dimensional system. But ν can arise as a VESD of a finite-dimensional system, which allows Theorem 3.6 to apply to (large) finite-dimensional linear algebra computations.

Also, as in Remark 2.5, we can see, from Theorem 3.6, that to obtain the fluctuations of quantities related to the numerical algorithms, it suffices to focus on the matrix $P(z)$ either at $z=0$ or $z=\infty$. In Section 4, we will focus on the spiked sample covariance matrix model and study these fluctuations, that is, the limiting behavior of $P(z)$.

4 Case Study: Spiked Sample Covariance Matrix Model

In this section, we focus our discussion on a concrete random matrix model, the celebrated spiked sample covariance matrix model, to illustrate how to conduct the analysis. Motivated by the applications in applied mathematics, we focus on the analysis of its limiting VESD; see Section 4.3 for more details. For any probability measure μ, its Stieltjes transform is defined as

$$m_\mu(z) = 2\pi \mathrm{i} c_0(z;\mu) = \int \frac{1}{x-z}\mu(\mathrm{d}x),\ z \in \mathbb{C}_+.$$

4.1 The deformed Marchenko–Pastur law

We first introduce the celebrated deformed Marchenko–Pastur (MP) law. Let X be an $N\times M$ random matrix with independent and identically distributed (iid) centered entries with variance M^{-1} and the population covariance matrix Σ_0 be a positive definite deterministic matrix satisfying some regularity conditions (cf. Assumption 2). Denote the sample covariance matrix and its companion, which are both random matrices, as follows

$$\mathcal{Q}_1 = \Sigma_0^{1/2} X X^* \Sigma_0^{1/2},\ \mathcal{Q}_2 = X^* \Sigma_0 X. \tag{4.1}$$

In the sequel, we assume that for some small constant $0<\tau<1$

$$\tau \le c_N := \frac{N}{M} \le \tau^{-1}. \tag{4.2}$$

Denote the spectral decomposition of Σ_0 as

$$\Sigma_0 = \sum_{k=1}^{N} \sigma_i \mathbf{v}_i \mathbf{v}_i^*, \quad 0 < \sigma_N \le \sigma_{N-1} \le \cdots \le \sigma_1 < \infty.$$

The Stieltjes transform $m(z)$ of the deformed MP law can be characterized as the unique solution of the following equation [51, Lemma 2.2]

$$z = f(m), \ \operatorname{Im} m(z) \ge 0,$$

where $f(x)$ is defined as

$$f(x) = -\frac{1}{x} + \frac{1}{M} \sum_{k=1}^{N} \frac{1}{x + \sigma_k^{-1}}. \tag{4.3}$$

Denote $\varrho = \varrho_{\Sigma_0,N}$ as the probability measure associated with m. Then ϱ is referred to as the *deformed MP law*, whose properties are summarized as follows; see Lemmas 2.5 and 2.6 of [51] for more details.

Lemma 4.1. The support of ϱ is a union of connected components on $\mathbb{R}_+$:

$$\operatorname{supp} \varrho = \bigcup_{k=1}^{q} [\mathsf{e}_{2k}, \mathsf{e}_{2k-1}] \subset (0, \infty), \tag{4.4}$$

where q depends on the ESD of Σ_0. Here $\mathsf{e}_1 \ge \mathsf{e}_2 \ge \cdots \ge \mathsf{e}_{2q}$ can be characterized as follows: there exists a real sequence $\{\mathsf{t}_k\}_{k=1}^{2q}$ such that $(x, m) = (\mathsf{e}_k, \mathsf{t}_k)$ are real solutions to the equations

$$x = f(m), \text{ and } f'(m) = 0.$$

Based on Lemma 4.1, we shall call the sequence of $\mathsf{e}_k, k = 1, 2, \cdots, 2q$, as the edges of the deformed MP law ϱ. To avoid repetition, we summarize the assumptions the will be used in the current paper. These assumptions are standard and commonly used in the random matrix theory literature; see Definition 2.7 of [51] for more details.

Assumption 2. We assume that (4.2) holds and $|c_N - 1| \ge \tau$. Moreover, for $X = (X_{ij})$, we assume that $X_{ij}, 1 \le i \le N, \ 1 \le j \le M$, are iid random variables such that

$$\mathbb{E}X_{ij} = 0, \ \mathbb{E}X_{ij}^2 = \frac{1}{M}.$$

Moreover, we assume that for all $k \in \mathbb{N}$, there exists some constant C_k such that

$$\mathbb{E}|\sqrt{M}X_{ij}|^k \leq C_k. \tag{4.5}$$

For Σ_0, we assume that for some small constant $0 < \tau_1 < 1$, the following holds:

$$\tau_1 \leq \sigma_N \leq \sigma_{N-1} \leq \cdots \leq \sigma_1 \leq \tau_1^{-1}.$$

Additionally, for the two sequences of $\{\mathsf{e}_k\}$ and $\{\mathsf{t}_k\}$ in Lemma 4.1, we assume that

$$\mathsf{e}_k \geq \tau_1, \ \min_{l \neq k} |\mathsf{e}_k - \mathsf{e}_l| \geq \tau_1, \ \min_i |\sigma_i^{-1} + \mathsf{t}_k| \geq \tau_1.$$

Finally, for any fixed small constant τ_2, there exists some constant $\varsigma = \varsigma_{\tau_1,\tau_2} > 0$ such that the density of ϱ in $[\mathsf{e}_{2k} + \tau_2, \mathsf{e}_{2k-1} - \tau_2]$ is bounded from below by ς.

Remark 4.1. We make a remark on the deformed MP law. Even though we will not study ϱ and its perturbation (i.e., the empirical spectral distribution (ESD)), we point out that ϱ satisfies Assumption 1. According to [51, Section A.2] (or Lemma 3.6 of [34], or Proposition 2.6 of [40]), under Assumption 2, we have that $\varrho(x) \sim \sqrt{\mathsf{e}_k - x}$, $x \in [\mathsf{e}_k - \tau, \mathsf{e}_k]$ for some small constant $\tau > 0$. Consequently, we can conclude that ϱ satisfies (2.16) by setting $\mathsf{a}_j = \mathsf{e}_{2j}, \mathsf{b}_j = \mathsf{e}_{2j-1}$ and $w_j = 0$. Moreover, (2)–(4) of Assumption 1 are satisfied due to Assumption 2.

4.2 The spiked model

We are now ready to state our model by adding r spikes to Σ_0, where $r \geq 0$ is some fixed integer. Let Σ be a spiked sample covariance matrix based on Σ_0 so that it admits the following spectral decomposition

$$\Sigma = \sum_{i=1}^{M} \widetilde{\sigma}_i \mathbf{v}_i \mathbf{v}_i^*,$$

where $\widetilde{\sigma}_i = (1 + d_i)\sigma_i$ such that $d_i > 0, i \leq r$ and $d_i = 0, i > r$. To ease our discussion, we assume the spikes are supercritical as summarized below following [32].

Assumption 3. For $i \leq r$, we assume that there exists some constant ϖ such that

$$\widetilde{\sigma}_i > -\mathsf{t}_1^{-1} + \varpi. \tag{4.6}$$

We also assume that $\sigma_i, 1 \leq i \leq r$ are distinct and bounded.

Then, the spiked sample covariance matrix and its companion are defined, respectively, as follows:

$$\widetilde{\mathcal{Q}}_1 := \Sigma^{1/2} X X^* \Sigma^{1/2}, \ \widetilde{\mathcal{Q}}_2 := X^* \Sigma X. \tag{4.7}$$

The above model is a generalization of Johnstone's spiked sample covariance matrix model [48]. Let $\{\lambda_i(\widetilde{\mathcal{Q}}_1)\}$ be the eigenvalues $\widetilde{\mathcal{Q}}_1$ in the decreasing order and $\{\widetilde{\mathbf{u}}_i\}$ be the associated eigenvector.

Under Assumption 3, we have the following result [32, Theorem 3.6]. Recall $f(x)$ in (4.3).

Lemma 4.2. Suppose Assumptions 2 and 3 hold. Then we have that for all $1 \leq i \leq r$,

$$\left|\lambda_i(\widetilde{\mathcal{Q}}_1) - f(-\widetilde{\sigma}_i^{-1})\right| = \mathrm{O}_{\mathbb{P}}(N^{-1/2}), \quad \text{and} \quad \left|\langle \widetilde{\mathbf{u}}_i, \mathbf{v}_i\rangle^2 - \frac{1}{\widetilde{\sigma}_i} \frac{f'(-\widetilde{\sigma}_i^{-1})}{f(-\widetilde{\sigma}_i^{-1})}\right| = \mathrm{O}_{\mathbb{P}}(N^{-1/2}).$$

4.3 VESDs and their limits

In this subsection, we introduce the VESDs and their deterministic limits. To be consistent with the notation of Section 2, we denote the VESDs as $\nu = \nu_N, \widetilde{\nu} = \widetilde{\nu}_N$ and their deterministic limits as $\mu = \mu_N, \widetilde{\mu} = \widetilde{\mu}_N$ for the non-spiked model in (4.1) and spiked model in (4.7), respectively.

For any projection, $\mathbf{b}$, we denote the VESD of $\mathcal{Q}_1$ as

$$\nu = \sum_{i=1}^N |\langle \mathbf{u}_i, \mathbf{b}\rangle|^2 \delta_{\lambda_i(\mathcal{Q}_1)}, \tag{4.8}$$

where $\{\mathbf{u}_i\}$ are the eigenvectors of $\mathcal{Q}_1$ and $\{\lambda_i(\mathcal{Q}_1)\}$ are its eigenvalues. Similarly, we denote the VESD of $\widetilde{\mathcal{Q}}_1$ as

$$\widetilde{\nu} = \sum_{i=1}^N |\langle \widetilde{\mathbf{u}}_i, \mathbf{b}\rangle|^2 \delta_{\lambda_i(\widetilde{\mathcal{Q}}_1)}.$$

The limits of ν and $\widetilde{\nu}$ can be characterized by the so-called anisotropic local law (cf. Lemmas C.1 and C.9). Especially, the Stieltjes transforms of μ and $\widetilde{\mu}$ can be characterized, respectively, as [33, 51]

$$m_\mu(z) = -\frac{1}{z}\mathbf{b}^*(1 + m(z)\Sigma_0)^{-1}\mathbf{b}, \ m_{\widetilde{\mu}}(z) = \sum_{i=1}^N \frac{\omega_i^2}{1+d_i}\left(-\frac{1}{z}(1+m(z)\sigma_i)^{-1} - \mathcal{L}_i\right), \tag{4.9}$$

where we denote

$$\omega_i = \mathbf{v}_i^* \mathbf{b}, \quad \mathcal{L}_i = \mathbf{1}(i \le r) z^{-1} (1 + m(z)\sigma_i)^{-2} (d_i^{-1} + 1 - (1 + m(z)\sigma_i)^{-1})^{-1}, \tag{4.10}$$

and recall that $m(z)$ is the Stieltjes transform of the deformed MP law.

Before concluding this subsection, we explain how the measures μ and $\widetilde{\mu}$ satisfy Assumption 1. First, using the inversion formula that $\mu\{[a,b]\} = \pi^{-1}\int_a^b \operatorname{Im} m_\mu(x+\mathrm{i}0^+)\mathrm{d}x$, it is easy to see from (4.9) that the density of μ, denoted as $\varrho_{\mathbf{b}}$ satisfies (see (3.4) of [33])

$$\varrho_{\mathbf{b}}(x) = \frac{\varrho(x)}{x} \mathbf{b}^* \Sigma_0 \big[I + 2\operatorname{Re} m(x + \mathrm{i}0^+)\Sigma_0 + |m(x+\mathrm{i}0^+)|^2 \Sigma_0^2 \big]^{-1} \mathbf{b}. \tag{4.11}$$

Under Assumption 2, it is easy to see that $\varrho_{\mathbf{b}}(x) \sim \varrho(x)$ so that μ satisfies Assumption 1 as discussed in Remark 4.1.

For the spiked model, it depends crucially on $\mathbf{b}$. We will need the following assumption to match the condition (5) of Assumption 1.

Assumption 4. For ω_i defined in (4.10) and all $1 \le i \le r$, we assume that either of the following holds:

$$\omega_i = 0, \text{ or } 1/D \le |\omega_i| \le D.$$

Under Assumption 4, on one hand, $\omega_i = 0$ for all $1 \le i \le r$, it is easy to see that μ and $\widetilde{\mu}$ coincide so that Assumption 1 holds. On the other hand, if some of ω_i are nonzero satisfying Assumption 4, without loss of generality, say only $\omega_1 \asymp 1$. Using the relation that $d_i^{-1} + 1 - (1 + m(f(-\widetilde{\sigma}_i^{-1}))\sigma_i)^{-1} = 0$, according to (4.9), Lemma 4.2 and Assumption 4, we find that $\widetilde{\mu}$ satisfies (2.16) by setting $\mathsf{a}_j = \mathsf{e}_{2j}, \mathsf{b}_j = \mathsf{e}_{2j-1}$ and $\mathsf{c}_1 = f(-\widetilde{\sigma}_1^{-1}), w_1 = \frac{1}{\widetilde{\sigma}_1} \frac{f'(-\widetilde{\sigma}_1^{-1})}{f(-\widetilde{\sigma}_1^{-1})}, p = 1$. The general setting can be analyzed similarly.

4.4 A general CLT

As we can see from Section 3.3, it suffices to establish the CLT of the following form,

$$\mathcal{Y} := \sqrt{M\eta} \oint_\Gamma g(z) c_0(z; \mu - \nu) \mathrm{d}z, \text{ or } \widetilde{\mathcal{Y}} := \sqrt{M\eta} \oint_\Gamma g(z) c_0(z; \widetilde{\mu} - \widetilde{\nu}) \mathrm{d}z, \tag{4.12}$$

where $g(z)$ is an analytic in a neighborhood of Γ and $\eta = \eta(n)$ depending on some other parameter n is as in (2.22). Here we recall again that n can be the order of orthogonal polynomials or the number of iterations in the numerical algorithms.

According to our applications, by Lemma 4.2 and the local law (cf. Lemma C.1), $g(z)$ can be purely deterministic and given by the entries of $M_n(z;\mu)/z^k, k=0,1$ after some proper normalization so that $\oint_\Gamma |g(z)||\mathrm{d}z| \asymp 1$ and $\mathcal{Y}$ is a real-valued random variable as required. The main results are reported in Theorem 4.3. We first introduce the following definition.

Definition 1. For two sequences of random vectors $\mathbf{x}_N, \mathbf{y}_N \in \mathbb{R}^k, N \geq 1$, we say they are asymptotically equal in distribution, denoted as $\mathbf{x}_N \simeq \mathbf{y}_N$, if they are tight and satisfy

$$\lim_{N\to\infty} \big(\mathbb{E}l(\mathbf{x}_N) - \mathbb{E}l(\mathbf{y}_N)\big) = 0,$$

for any bounded continuous function $l : \mathbb{R}^k \to \mathbb{R}$.

Then we provide some notation. Denote

$$\Pi_1(z) := -\frac{1}{z}(1 + m(z)\Sigma_0)^{-1}, \tag{4.13}$$

and for any deterministic vectors $\mathbf{h}_1, \mathbf{h}_2 \in \mathbb{R}^N$, we define

$$\mathsf{V}_1(\mathbf{h}_1, \mathbf{h}_2) := \frac{\eta}{2\pi^2} \oint_\Gamma \oint_\Gamma \sqrt{z_1 z_2} g(z_1) g(z_2) \big[\mathbf{h}_1^*(1 + m(z_1)\Sigma_0)^{-1}\Sigma_0 \Pi_1(z_2)\mathbf{h}_2\big] \tag{4.14}$$

$$\times \left[\frac{\mathbf{h}_1^*(\Pi_1(z_1) - \Pi_1(z_2))\mathbf{h}_2}{z_1 - z_2}\right] \mathrm{d}z_1 \mathrm{d}z_2, \tag{4.15}$$

where we used the convention that

$$\lim_{z_1\to z_2} \frac{\mathbf{h}_1^*(\Pi_1(z_1) - \Pi_1(z_2))\mathbf{h}_2}{z_1 - z_2} = \mathbf{h}_1^* \Pi_1'(z_1)\mathbf{h}_2,$$

and

$$\mathsf{V}_2(\mathbf{h}_1, \mathbf{h}_2) := -\frac{\eta}{4\pi^2}\Big(\oint_\Gamma \oint_\Gamma g(z_1)g(z_2) z_1 z_2 m(z_1) m(z_2) \mathcal{K}(z_1, z_2) \mathrm{d}z_1 \mathrm{d}z_2\Big), \tag{4.16}$$

where $\mathcal{K}(z_1, z_2)$ is defined by

$$\mathcal{K}(z_1, z_2) := \sqrt{z_1} \sum_i (\Sigma_0^{1/2} \Pi_1(z_1) \mathbf{h}_1 \mathbf{h}_2^* \Pi_1(z_1) \Sigma_0^{1/2})_{ii} (\Sigma_0^{1/2} \Pi_1(z_2) \mathbf{h}_1 \mathbf{h}_2^* \Pi_1(z_2))_{ii}.$$

Let κ_4 be the cumulant of the random variable X_{ij} as defined in (C.10).

Theorem 4.3. Suppose $\mathcal{Y}$ and $\widetilde{\mathcal{Y}}$ are real valued. Suppose that Assumption 2 holds, then we have that

$$\mathcal{Y} \simeq \mathcal{N}(0, \mathsf{V}_1(\mathbf{b}, \mathbf{b}) + \kappa_4 \mathsf{V}_2(\mathbf{b}, \mathbf{b})).$$

Moreover, if Assumptions 3 and 4 hold,

$$\widetilde{\mathcal{Y}} \simeq \mathcal{N}(0, \widetilde{\mathsf{V}}_1 + \kappa_4 \widetilde{\mathsf{V}}_2),$$

where we used the notation that for $k = 1, 2$,

$$\widetilde{\mathsf{V}}_k := \sum_{i=1}^{N} \frac{\omega_i^2}{1+d_i} \left(\mathsf{V}_k(\mathbf{v}_i, \mathbf{v}_i) - \mathsf{V}_k(\mathbf{l}_i, \mathbf{v}_i) - \mathsf{V}_k(\mathbf{v}_i, \mathbf{l}_i) - \mathsf{V}_k(z^{-1}\mathbf{l}_i, \mathbf{l}_i) \right).$$

where ω_i is defined in (4.10) and $\mathbf{l}_i$ is defined in (C.68) after some additional necessary notation is introduced.

Proof. See Appendix C. ■

Remark 4.2. In our applications, when the deterministic function $g(z)$ is properly normalized, it is easy to check that

$$\mathsf{V}_1(\mathbf{b}, \mathbf{b}) \asymp 1, \ \mathsf{V}_2(\mathbf{b}, \mathbf{b}) \asymp \eta.$$

We recall from (2.22) that $\mathsf{V}_2 = \mathrm{O}(n^{-2})$. Consequently, when n diverges with any polynomial order, $\mathsf{V}_2(\mathbf{b}, \mathbf{b})$ can be negligible asymptotically and hence the fluctuations only depend on the first two moments and is therefore more universal. Similar phenomenon has been observed in the mesoscopic CLT of random matrix theory, see, for example, [6, 15, 45, 59, 76].

Remark 4.3. We provide a few examples to illustrate the results of the spiked model, that is, the CLT of $\widetilde{\mathcal{Y}}$. As can be seen in (C.68),

$$\mathbf{l}_i = \mathbf{0}, \ i > r.$$

Consequently, if $\mathbf{b} \in \mathrm{Span}(\{\mathbf{v}_i\}_{i>r})$, we find that

$$\widetilde{\mathsf{V}}_k = \mathsf{V}_k(\mathbf{b}, \mathbf{b}), \ k = 1, 2.$$

That is to say, when $\mathbf{b}$ lies in the orthogonal complement of the spiked eigenvectors, the distribution is the same with the non-spiked model. Moreover, when $\mathbf{b} = \mathbf{v}_{i_*}$, $1 \le i_* \le r$, we have that

$$\mathbf{l}_{i_*} = \frac{-1 - m(z)\sigma_{i_*}}{d_{i_*} + 1 - (1 + m(z)\sigma_{i_*})^{-1}} \mathbf{b}.$$

Consequently, we can simplify $\widetilde{\mathsf{V}}_k$ to

$$\widetilde{\mathsf{V}}_k := (1 + d_{i_*})^{-1} \big(\mathsf{V}_k(\mathbf{b}_{i_*}, \mathbf{b}_{i_*}) - \mathsf{V}_k(\mathbf{l}_{i_*}, \mathbf{b}_{i_*}) - \mathsf{V}_k(\mathbf{b}_{i_*}, \mathbf{l}_{i_*}) - \mathsf{V}_k(z^{-1}\mathbf{l}_{i_*}, \mathbf{l}_{i_*})\big).$$

As a consequence of Theorem 4.3, we can establish the asymptotic fluctuations of the associated orthogonal polynomials.

Corollary 4.4. Suppose the assumptions of Theorem 4.3 hold. Let the parameters $\mathfrak{c}, \mathfrak{g}, \mathsf{G}$ and $\{\mathsf{c}_j\}$ in (2.23), Θ in (A.2) and $\gamma(z)$ in (A.3) defined by the limiting VESD as in (4.11). Denote

$$\mathsf{L} := \mathfrak{c}^{(p-n)} \, \mathrm{e}^{(n-p)\mathfrak{g}(z) + \mathsf{G}(z) - \mathsf{G}(\infty)} \times \left[\prod_{j=1}^{p} (z - \mathsf{c}_j) \right],$$

$$\mathsf{E}_1 := \frac{\left(\frac{\gamma(z)+\gamma(z)^{-1}}{2}\right) \Theta_1(z; \mathbf{d}_2; (n-p)\boldsymbol{\Delta} + \boldsymbol{\zeta})}{\Theta_1(\infty; \mathbf{d}_2; (n-p)\boldsymbol{\Delta} + \boldsymbol{\zeta})}, \quad \mathsf{E}_2 := \mathrm{e}^{2\mathsf{G}(\infty)} \frac{\left(\frac{\gamma(z)^{-1}-\gamma(z)}{2\mathrm{i}}\right) \Theta_1(z; \mathbf{d}_1; (n-p)\boldsymbol{\Delta} + \boldsymbol{\zeta})}{\Theta_2(\infty; \mathbf{d}_1; (n-p)\boldsymbol{\Delta} + \boldsymbol{\zeta})}.$$

For the non-spiked model, when $C \log N \le n \le N^{1/6-\epsilon}$ for some $C > 0$ sufficiently large and $\epsilon > 0$, sufficiently small, for $z \in \mathbb{R} \setminus \mathrm{supp}(\mu)$, we have

$$\frac{\sqrt{M}}{n\mathsf{C}_g} \left(\pi_n(z; \nu) - \mathsf{L}\mathsf{E}_1\right) \simeq \mathcal{N}\left(0, \frac{\mathsf{L}}{\mathsf{C}_g^2} (\mathsf{V}_1(\mathbf{b}, \mathbf{b}) + \kappa_4 \mathsf{V}_2(\mathbf{b}, \mathbf{b}))\right),$$

where V_1 and V_2 are defined as in (4.14) and (4.16) by letting $\eta = n^{-2}$ and

$$g(z') = \frac{1}{2\pi\mathrm{i}} \frac{1}{z' - z} \left(\mathsf{E}_1 M_n(z')_{11} + \mathsf{E}_2 M_n(z')_{12}\right), \; \mathsf{C}_g = \oint_{\Gamma} |g(z')| |\mathrm{d}z'|,$$

where M_n is defined in (2.21). Similar results hold for the spiked model.

Proof. The proof follows directly from Theorem 4.3 and (2.23). ■

Remark 4.4. The normalization is used to ensure that, $\oint_\Gamma |g(z')|/\mathsf{C}_g|\mathrm{d}z'|$ is bounded from below and above so that Theorem 4.3 applies. Using an analogous discussion, we can derive the CLT for the Cauchy transforms as in (2.23). Since the concerned quantities of the numerical algorithms depend on the orthogonal polynomials and their Cauchy transforms, we can also obtain the asymptotic fluctuation of these algorithms using Theorems 3.6 and 4.3. We omit further details here.

5 Numerical Simulations and Some Discussions

We now provide numerical simulations of our estimates and perturbation theory to demonstrate the asymptotic behavior of both matrix factorizations and iterative algorithms (cf. Section 3.1) applied to the spiked sample covariance model.

5.1 Calculations of key parameters

As we have seen in the results of Sections 2 and 3 that many essential parameters need to be estimated before the application of Theorem 3.6. The first quantity is the density of the limiting VESD, its support, and strength of the spikes. In Appendix D, we provide a numerical method to approximate this. The outline of the procedure is:

- First, to compute the asymptotic support of the measure we use a rootfinder guided by Lemma 4.1.
- Second, to compute the asymptotic location of the spikes we use Lemma 4.2.
- Then we fit the coefficients in a mapped Chebyshev approximation of the density h_j on $[\mathsf{a}_j, \mathsf{b}_j]$ by solving a constrained optimization problem.

The method works with the empirical resolvent $\langle \mathbf{b}, (W - z)^{-1}\mathbf{b}\rangle$ or with the limiting Stieltjes transforms m_μ or $m_{\tilde{\mu}}$. In the former case, one should average over a number of trials. With the density function approximated in a useable form, we can calculate the other parameters. Appendix D outlines how to then approximate, with good accuracy, the limiting Jacobi matrix $\mathcal{J}(\mu)$ from which many other quantities of interest are easily computable.

Below, we also compute $\mathfrak{g}(0)$. Since g in Assumption 1 is small in our computations one can directly implement the procedure outlined in Section A.2.2 using the methodology of [73, Section 11.6.1] to compute the integrals that arise.

5.2 Performance of CGA with random inputs

In this subsection, we work on CGA when W is a spiked sample covariance matrix. We first work on an example where the support of the limiting VESD consists of two disjoint intervals (i.e., a single gap) with spikes. Then we study a three intervals (i.e., two gaps) case. Finally, we study the halting time of CGA, that is, the number of iterations needed before CGA terminates according to some stopping rule.

In the computations that follow, it is interesting to compare what results to the classical Chebyshev upper bound for the convergence of CGA [46]:

$$\frac{\|\mathbf{r}_n\|_2}{\|\mathbf{r}_{n-1}\|_2} \le \delta_{\mathrm{Cheb}}^{-1}, \quad \delta_{\mathrm{Cheb}} := \frac{\sqrt{\kappa(W)}+1}{\sqrt{\kappa(W)}-1}, \quad \kappa(W) = \lambda_{\max}(W)/\lambda_{\min}(W). \tag{5.1}$$

5.2.1 *CGA: single gap with spikes*

Consider the spiked sample covariance matrix $W = \Sigma^{1/2} X X^* \Sigma^{1/2}$ where X is $N \times M$, has iid entries,

$$\begin{aligned} &\Sigma_0 = \mathrm{diag}(8I, I), \quad X_{ij} \overset{\mathscr{L}}{=} \mathcal{N}(0, M^{-1}), \\ &M = \lfloor N/0.3 \rfloor, \quad d_1 = 1, \quad d_2 = 0.5, \quad d_i = 0,\ i \ge 3, \end{aligned} \tag{5.2}$$

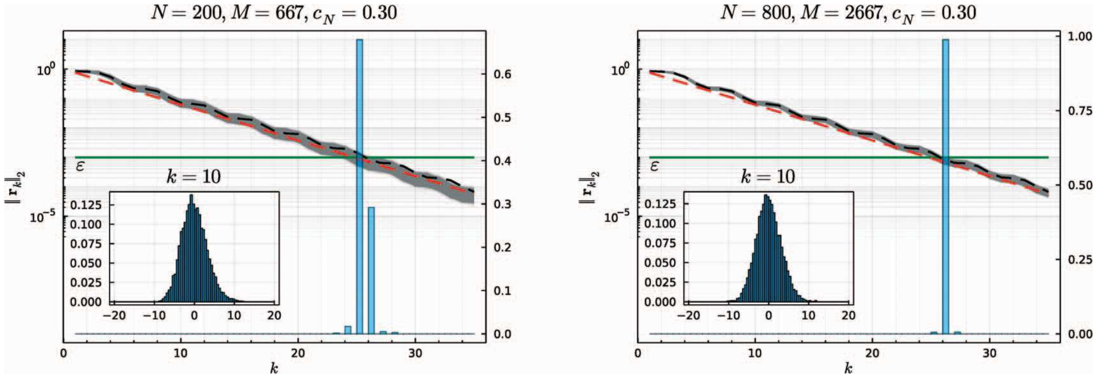
and I is the $N/2 \times N/2$ identity matrix. We choose $X_{ij} \overset{\mathscr{L}}{=} \mathcal{N}(0, M^{-1})$ for convenience because, as we have shown, the same limiting behavior will happen for any other admissible entry distribution. In Figure 1, we apply the CGA to $W\mathbf{x} = \mathbf{b}$ with $2\mathbf{b} = \mathbf{f}_1 + \mathbf{f}_2 + \mathbf{f}_3 + \mathbf{f}_N$. The residuals encountered at iteration k concentrate on the black dashed curve that is computed utilizing the results of Section 3 with parameters calculated using methodology outlined in Section 5.1. In particular, the choices of the parameters are

$$\begin{aligned} &\mathtt{a}_1 \approx 0.279, \quad \mathtt{b}_1 \approx 1.667, \quad \mathtt{a}_2 \approx 3.192, \quad \mathtt{b}_2 \approx 15.562, \\ &\delta := \mathrm{e}^{\mathfrak{g}(0)} \approx 1.322, \\ &\mathtt{c}_1 \approx 20.319, \quad \mathtt{c}_2 \approx 33.755. \end{aligned} \tag{5.3}$$

In addition to the black curve, we also provide a red curve. The motivation is as follows. According to Corollary 3.3 and Theorem 3.6 (or Remark 3.2), we find that

$$\frac{\|\mathbf{r}_n\|_2}{\|\mathbf{r}_{n-1}\|_2} \approx \mathrm{e}^{-\mathfrak{g}(0)}.$$

Note that in this case $\delta_{\mathrm{Cheb}} \approx 1.309$ indicating that when one accounts for the gap in the spectrum, a faster convergence rate is predicted.



(a) (b)

Fig. 1. The CGA runs on the single gap matrix in (5.2). The black oscillatory dashed curve indicates the large N limit for the residual norms $\|\mathbf{r}_k\|_2$ at step k. The shaded gray area is an ensemble of 10000 runs of the conjugate gradient algorithm, displaying the residuals that resulted. The red dashed line is given by δ^{-k}, $\delta = \mathrm{e}^{\mathfrak{g}(0)}$. The overlaid histogram shows the rescaled fluctuations in the norm of the residual at $k = 10$. As $N \to \infty$, this approaches a Gaussian density. Lastly, the histogram in the main frame gives the halting distribution $\tau(W, \mathbf{b}, \epsilon) = \min\{k : \|\mathbf{r}_k\|_2 < \epsilon\}$ for $\epsilon = 10^{-3}$ (green horizontal line), that is, the statistics of the number of iterations required to achieve $\|\mathbf{r}_k\|_2 < \epsilon$. We can see that our results in Theorem 3.6 are reasonably good even for $N = 200$. The accuracy improves when N increases.

Then after being properly scaled, we can use $\mathrm{e}^{-n\mathfrak{g}(0)}$ for the prediction. We find that both our black and red curves are accurate even for small values of N. Furthermore, a remarkable feature of (3.8) and (3.9) is that the random components are contained in the P_{11} and P_{12} terms, which have a common exponential factor. This implies that the fluctuations are on the same exponential scale as the asymptotic mean. We demonstrate this by considering $\|\mathbf{r}_k\|_2\delta^k$ in Figure 2.

5.2.2 *CGA: Two gaps*

Consider the non-spiked sample covariance matrix $W = \Sigma_0^{1/2} X X^* \Sigma_0^{1/2}$ where X is $N \times M$, has iid entries,

$$\Sigma_0 = \mathrm{diag}(3.8I, 1.2I, 0.25I), \quad X_{ij} \stackrel{\mathscr{L}}{=} \mathcal{N}(0, M^{-1}), \quad M = \lfloor N/0.3 \rfloor, \tag{5.4}$$

and I is the $N/3 \times N/3$ identity matrix (We choose I here to be either $\lfloor N/3 \rfloor \times \lfloor N/3 \rfloor$ or $\lceil N/3 \rceil \times \lceil N/3 \rceil$.). In Figure 3, we apply the CGA to $W\mathbf{x} = \mathbf{b}$ with $\sqrt{3}\mathbf{b} = \mathbf{f}_1 + \mathbf{f}_{N/2} + \mathbf{f}_N$.

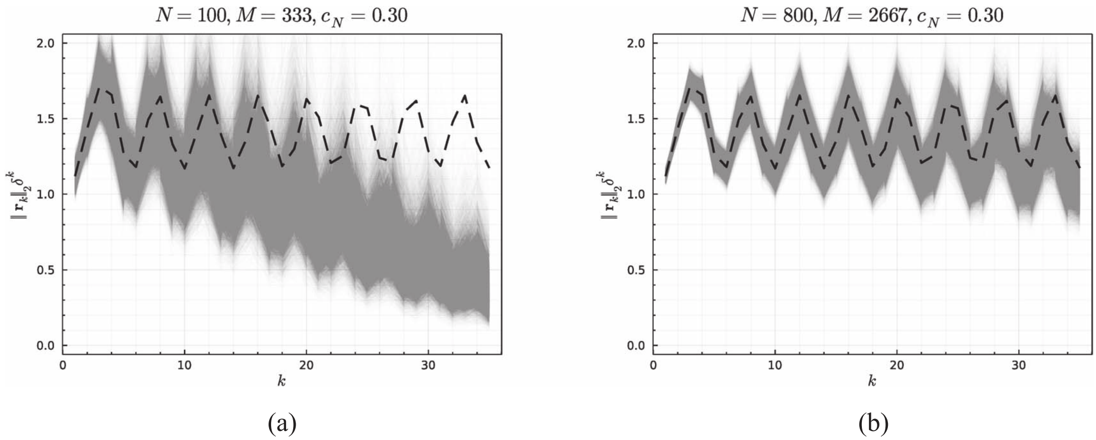


Fig. 2. The CGA runs on the single gap matrix in (5.2). The black oscillatory dashed curve indicates the large N limit for the scaled residual norms $\|\mathbf{r}_k\|_2\delta^k$ at step k. The shaded gray area is an ensemble of 10000 runs of the conjugate gradient algorithm, displaying the scaled residuals that resulted. We can see that our predictions in Theorem 3.6 are quite accurate once N is reasonably large.

We again report both the black and red curves, and they are reasonably accurate. The choices of the parameters we find are

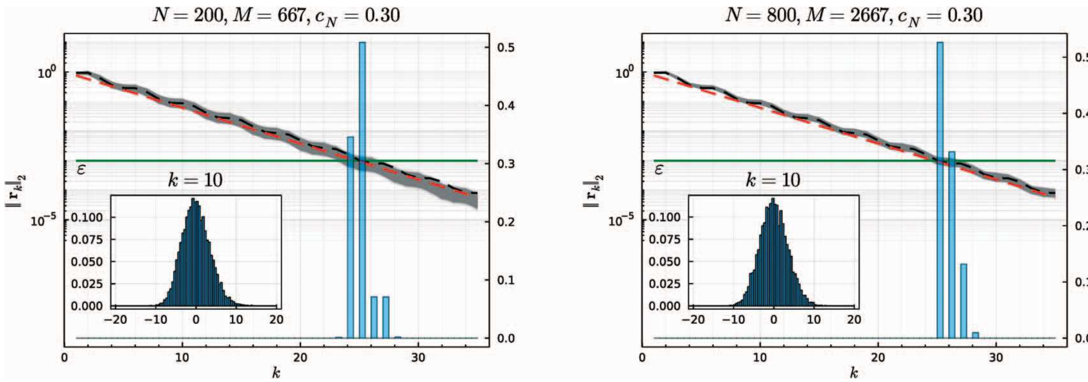
$$\begin{aligned} \mathrm{a}_1 &\approx 0.080, \quad \mathrm{b}_1 \approx 0.349, \\ \mathrm{a}_2 &\approx 0.496, \quad \mathrm{b}_2 \approx 1.828, \\ \mathrm{a}_3 &\approx 2.029, \quad \mathrm{b}_3 \approx 6.767, \\ \delta &:= \mathrm{e}^{\mathfrak{g}(0)} \approx 1.248. \end{aligned}$$

Note that in this case $\delta_{\mathrm{Cheb}} \approx 1.244$.

5.3 The Jacobi and Cholesky matrices

In this subsection, we analyze the entries of the Jacobi matrix in (3.5) and its associated Cholesky decomposition in (3.6). We first pause to review some classical results. The Householder tridiagonalization of a real symmetric or complex Hermitian matrix W is a fundamental numerical process. The process is succinctly described by the selection of a sequence of Householder reflectors, $U_1, \ldots, U_N$, so that

$$U_N U_{N-1} \cdots U_1 W U_1^* \cdots U_{N-1}^* U_N^* = J,$$



(a) (b)

Fig. 3. The CGA runs on the two gap matrix in (5.4). The details of the figures are similar to the captions of Figure 1. We can see that our results in Theorem 3.6 remain accurate for the new choice of Σ_0.

is a symmetric tridiagonal matrix. The typical convention is to select each U_j so that the only non-zero entry in the first row and first column is a one in the $(1,1)$-entry. The off-diagonal entries of J can be chosen to be non-negative.

When $W \stackrel{\mathscr{L}}{=} XX^*$, $X_{ij} \stackrel{\mathscr{L}}{=} \mathcal{N}(0, M^{-1})$, the case of a Wishart matrix, the distribution of J can be calculated explicitly [37, 69] and is given by

$$J \stackrel{\mathscr{L}}{=} LL^T, \quad L = \frac{1}{\sqrt{M}} \begin{bmatrix} \chi_{\beta M} & & & & \\ \chi_{\beta(N-1)} & \chi_{\beta(M-1)} & & & \\ & \chi_{\beta(N-2)} & \chi_{\beta(M-2)} & & \\ & & \ddots & \ddots & \\ & & & \chi_{\beta} & \chi_{\beta(M-N+1)} \end{bmatrix}, \tag{5.5}$$

where χ_γ is a χ-distributed random variable with γ degrees of freedom and all the entries of L are independent. Here $\beta = 1$ if the matrix W has real entries. In another way of speaking, the matrix L gives the distribution of the Cholesky factorization of the tridiagonalization of a Wishart matrix. One can generalize this tridiagonalization by asking that the first column of $U_1^* \cdots U_{N-1}^* U_N^*$ be a prescribed vector $\mathbf{b}$ so that

$$T = T(W, \mathbf{b}), \quad L = L(W, \mathbf{b}). \tag{5.6}$$

This can be accomplished by simply constructing a matrix U_0, $U_0^* U_0 = I$ whose first column is $\mathbf{b}$ and apply the Householder tridiagonalization procedure to $U_0^* W U_0$. In this case, the tridiagonal matrix that results coincides with the output of the Lanczos algorithm (In numerical linear algebra these two methods are treated as distinct, in part, because they have vastly different behavior in finite-precision arithmetic.) .

More is true. Consider the discrete measure

$$\nu = \nu_{W,\mathbf{b}} = \sum_{i=1}^{N} |\langle \mathbf{u}_i, \mathbf{b} \rangle|^2 \delta_{\lambda_i(W)},$$

for a general positive definite matrix W. Then,

$$T(W, \mathbf{b}) = \mathcal{J}_N(\nu),$$

$$L(W, \mathbf{b}) = \mathcal{L}_N(\nu),$$

which directly connects the output of the algorithms to the VESD. In the context of Wishart matrix, supposing $c_N = N/M \to \mathrm{c} \in (0, 1]$, one can immediately see that the (k, k) and $(k, k-1)$ entries of L in (5.5) tend to 1 and $\sqrt{\mathrm{c}}$, respectively, provided $k \ll N$. Furthermore, the fluctuations will be Gaussian, by the central limit theorem. It is of intrinsic interest to ask if this phenomenon persists for the spiked sample covariance model we analyze here. Our results establish this for $k \ll N^{1/6}$ and we conjecture it holds for $k \ll N$.

We now explain simulations based on the matrix model defined by (5.2) to demonstrate both our results and add evidence that that $k \ll N$ is necessary. Let ν be given by (4.8) with limiting measure μ using the setting (5.2). As stated, the tridiagonalization of the spiked sample covariance model and its Cholesky factorization are given by $J_N(\nu)$ and $\mathcal{L}_N(\nu)$ using the notation of (3.5) and (3.6). In this section, we examine $a_k(\nu)$ and $\alpha_k(\nu)$ for $k \leq 8N^{1/6} + 10$ and $k \leq N/3$ using the results of Section 3. Figure 4 demonstrates a consequence of our results (While our results technically only hold for $k \ll N^{1/6}$, allowing $k \leq 8N^{1/6} + 10$ demonstrates that we expect our results up to hold up to this threshold.) that the entries of $\mathcal{J}_k(\nu)$ concentrate on those of $\mathcal{J}_k(\mu)$ for $k \ll N^{1/6}$. If we allow k to be proportional to N, we do not expect this to occur as Figure 5 demonstrates.

Lastly, we consider the fluctuations of the diagonal elements of $\mathcal{J}(\nu)$ where ν is the VESD in (4.8). We have shown that as $N \to \infty$, for k fixed, the fluctuations of a_k are Gaussian. Furthermore, by Theorem 4.3, Corollary 4.4, and Remark 4.4, the variance depends on the fourth moment of the matrix entries. We confirm this clearly in the top

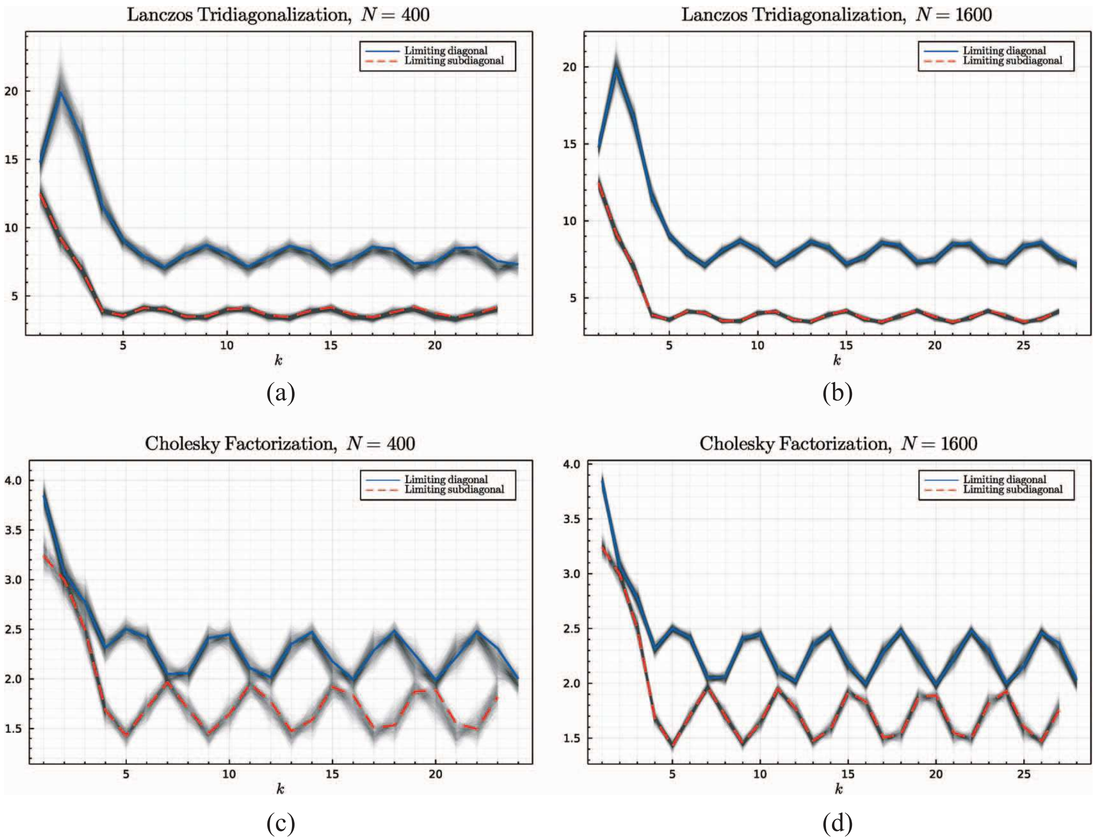


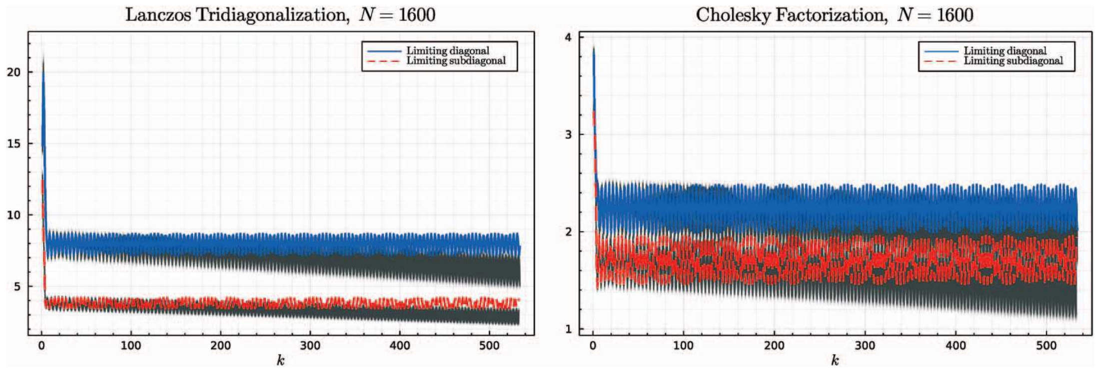
Fig. 4. The first k entries of the matrices $J_N(\nu)$ and $\mathcal{L}_N(\nu)$ for $k \leq 5N^{1/6} + 10$ in the case of (5.2). The solid blue and dashed red curves give the large N limit of the diagonal and subdiagonal, respectively, computed using the results of Theorem 3.6 with the parameters calculated using the methods outlined in Section 5.1. The shaded region is produced using 1000 samples for the displayed value of N. This demonstrates that if ν is given by (4.8) with limiting measure μ, using the setting (5.2), then the entries of $\mathcal{J}_k(\nu)$ concentrate on those of $\mathcal{J}_k(\mu)$ if $k \ll N^{1/6}$.

two panels of Figure 6. But as Remark 4.2 points out, as k increases, the dependence on the fourth moment should become negligible. Figure 6 demonstrates that this happens quickly.

Appendix A. Orthogonal Polynomials and Their Asymptotics: Proof of Theorem 2.2

A.1 A Riemann surface

In order to describe the asymptotics of polynomials orthogonal with respect to a measure μ from (2.16) satisfying the assumptions (1)–(5), we need to describe a Riemann surface.



(a) (b)

Fig. 5. The first k entries of the matrices $J_N(\nu)$ and $\mathcal{L}_N(\nu)$ for $k \le N/3$ in the case of (5.2). We can see that even though the entries of $\mathcal{J}_k(\nu)$ concentrate on those of $\mathcal{J}_k(\mu)$ for not-so-large k as in Figure 4, the prediction becomes inaccurate when k is larger.

General references for what follows are [4, 7, 21]. Associated with the intervals $[\mathfrak{a}_j, \mathfrak{b}_j]$, $1 \le j \le g+1$, is a Riemann surface, described by the solution set of

$$w^2 = \prod_{j=1}^{g+1} (z - \mathfrak{a}_j)(z - \mathfrak{b}_j) =: P_{2g+2}(z),$$

in $\mathbb{C}^2$. Consider a cut version of the complex plane:

$$\hat{\mathbb{C}} = \mathbb{C} \setminus \bigcup_{j=1}^{g+1} [\mathfrak{a}_j, \mathfrak{b}_j].$$

Then define a sectionally analytic function

$$R : \hat{\mathbb{C}} \to \mathbb{C}, \quad R(z)^2 = P_{2g+2}(z), \quad R(z) \to 1, \quad \text{as} \quad z \to \infty.$$

A Riemann surface Γ can be constructed by adjoining copies of $\hat{\mathbb{C}}$; see Figure A7 for an illustration and a description of the $\mathfrak{a}$-cycles and $\mathfrak{b}$-cycles. We have a natural projection $\pi : \Gamma \to \mathbb{C}$ defined by $\pi((z, w)) = z$ and its right-inverses $\pi_j^{-1}(z) = (z, (-1)^{j+1} R(z))$, $j = 1, 2$.

As is well-known (see [7], for example) a basis for holomorphic differentials on Γ is given by

$$\mathrm{d}\nu_j = \frac{z^{j-1}}{R(z)} \mathrm{d}z, \quad j = 1, 2, \ldots, g.$$

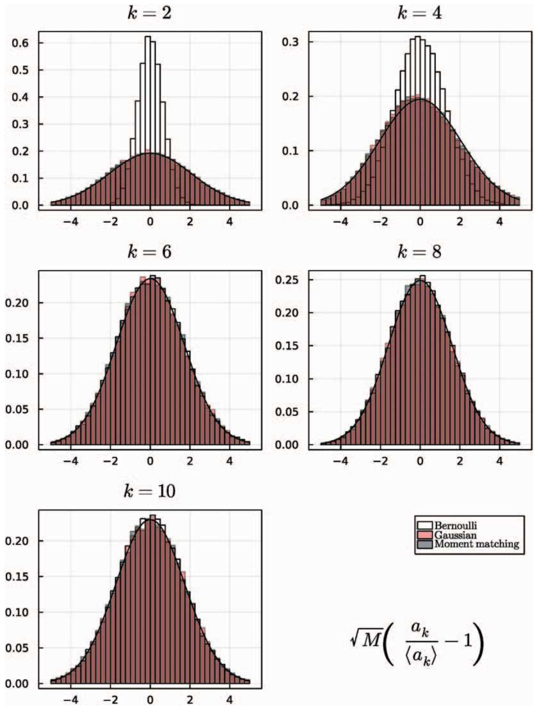


Fig. 6. Statistics of a_k for the model in (5.2) for different choices of distributions on the entries X_{ij} when $N = 1000$. For each choice of distribution, we plot a histogram for $\sqrt{M}(a_k/\langle a_k\rangle - 1)$ using 50,000 samples where $\langle\cdot\rangle$ gives the sample average over these 50,000 samples. The thin black curve is the density for a normal distribution with mean zero and variance determined by the sample variance of $\sqrt{M}(a_k/\langle a_k\rangle - 1)$ when $X_{ij} \overset{\mathscr{L}}{=} \mathcal{N}(0, M^{-1})$. The shaded red area gives the histogram for $\mathcal{N}(0, M^{-1})$ entries, the shaded gray area gives the histogram for the discrete distribution on $\{-1/\sqrt{M}, 0, 1/\sqrt{M}\}$ that matches its first four moments with $\mathcal{N}(0, M^{-1})$, and the white histogram is produced by $X_{ij} = \pm 1/\sqrt{M}$ with equal probability (Bernoulli). For smaller values of k the variance clearly is different between the moment matching distribution and the Bernoulli distribution. As k increases, this difference dramatically diminishes, as predicted in Corollary 4.4.

Define the $g \times g$ period matrix A by

$$A_{ij} = \oint_{\mathfrak{a}_i} \mathrm{d}\nu_j.$$

Note that if $c = \begin{bmatrix} c_1 & c_2 & \cdots & c_g \end{bmatrix}^T = A^{-1} e_j$ for the standard basis vector e_j, then

$$\oint_{\mathfrak{a}_i} \sum_{k=1}^{g} c_k \mathrm{d}\nu_k = \sum_{k=1}^{g} c_k A_{ik} = e_i^T A c = e_i^T e_j = \delta_{ij}.$$

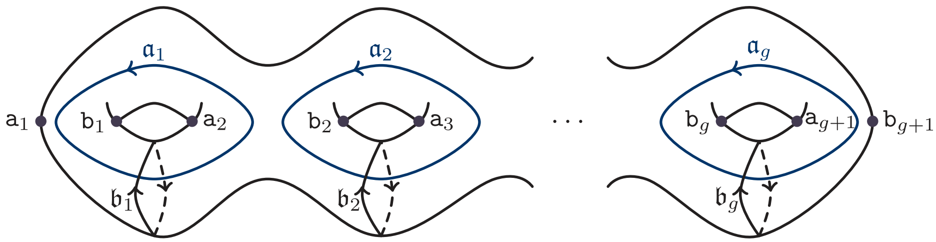


Fig. A7. An illustration of the Riemann surface Γ.

So, we define a basis of normalized differentials

$$\begin{bmatrix} d\omega_1 \\ d\omega_2 \\ \vdots \\ d\omega_g \end{bmatrix} = 2\pi i A^{-1} \begin{bmatrix} d\nu_1 \\ d\nu_2 \\ \vdots \\ d\nu_g \end{bmatrix},$$

which satisfies

$$\oint_{\mathfrak{a}_i} d\omega_j = 2\pi i \delta_{ij}.$$

The invertibility of the matrix A follows from abstract theory as in [13].

Now fix the base point $a = \mathtt{a}_1$ and define

$$u(z) = \left(\int_a^z d\omega_j \right)_{j=1}^{g}, \quad z \notin \mathbb{R},$$

where the path of integration is taken to be a straight line connecting a to z. Note that this extends to a vector-valued holomorphic function (We abuse notation here and treat u as both a function of $z \in \mathbb{C} \setminus \mathbb{R}$ and a function of $P \in \Gamma$.) $u(P)$ on the Riemann surface Γ provided Γ is cut along the cycles $\{\mathfrak{a}_1, \ldots, \mathfrak{a}_g, \mathfrak{b}_1, \ldots, \mathfrak{b}_g\}$, making it simply connected. Another important feature is that for $z \in \hat{\mathbb{C}}$, $u(\pi_1^{-1}(z)) = -u(\pi_2^{-1}(z))$.

Define the associated Riemann matrix of $\mathfrak{b}$ periods,

$$\tau = (\tau_{ij}) = \left(\int_{\mathfrak{b}_j} d\omega_i \right)_{1 \le i,j \le g}.$$

Note that τ is symmetric and pure imaginary and $-i\tau$ is positive definite. Next, define the vector $\mathbf{k}$ of Riemann constants component wise via

$$\mathbf{k}_j = \frac{2\pi i + \tau_{jj}}{2} - \frac{1}{2\pi i} \sum_{\ell \neq j} \oint_{\mathfrak{a}_\ell} u_j d\omega_\ell, \quad j = 1, 2, \ldots, g.$$

The associated theta function is given by

$$\theta(z;\tau) = \sum_{m\in\mathbb{Z}^g} \exp\left(\frac{1}{2}(m,\tau m) + (m,z)\right), \quad z \in \mathbb{C}^g, \tag{A.1}$$

where $(\cdot,\cdot)$ is the real scalar product. This series is convergent because τ has a negative-definite real part. The following hold:

$$\theta(z + 2\pi \mathrm{i} e_j;\tau) = \theta(z;\tau),$$
$$\theta(z + \tau e_j;\tau) = \exp\left(-\frac{1}{2}\tau_{jj} - z_k\right)\theta(z;\tau).$$

A divisor $D = \sum_j n_j P_j$ is a formal sum of points $\{P_j\}$ on the Riemann surface Γ. The Abel map of a divisor is defined via

$$\mathcal{A}(D) = \sum_j n_j u(P_j).$$

We now determine the jumps satisfied by the vector-valued function,

$$\Theta(z;d;v) = \Theta(z) := \begin{bmatrix} \frac{\theta(u(z)+v-d;\tau)}{\theta(u(z)-d;\tau)} & \frac{\theta(-u(z)+v-d;\tau)}{\theta(-u(z)-d;\tau)} \end{bmatrix}, \quad z \notin \mathbb{R}. \tag{A.2}$$

Note that the first component function is nothing more than $\frac{\theta(u(P)+v-d;\tau)}{\theta(u(P)-d;\tau)}$ restricted to the first sheet. The same is true for the second component function on the second sheet. The vector v is left arbitrary for now, and it will be chosen in a crucial way in what follows.

Then note that

$$u^+(z) + u^-(z) = \left(2\sum_{k=1}^{j-1}\int_{\mathrm{b}_k}^{\mathrm{a}_{k+1}} \mathrm{d}\omega_\ell\right)_{\ell=1}^{g} = \left(\sum_{k=1}^{j-1}\oint_{\mathfrak{a}_k} \mathrm{d}\omega_\ell\right)_{\ell=1}^{g} = 2\pi \mathrm{i}\mathsf{N}, \quad z \in [\mathrm{a}_j, \mathrm{b}_j],$$

for a vector N of zeros and ones. Then we compute

$$u^+(z) - u^-(z) = \left(2\sum_{k=1}^{j}\int_{\mathrm{a}_k}^{\mathrm{b}_k} \mathrm{d}\omega_\ell\right)_{\ell=1}^{g} = \left(\oint_{\mathfrak{b}_j} \mathrm{d}\omega_\ell\right)_{\ell=1}^{g} = \tau e_j, \quad z \in [\mathrm{b}_j, \mathrm{a}_{j+1}].$$

Then check

$$\frac{\theta\left(\pm u(z) + \tau e_j + v - d;\tau\right)}{\theta\left(\pm u(z) + \tau e_j - d;\tau\right)} = \mathrm{e}^{\pm v_k}\frac{\theta\left(\pm u(z) + v - d;\tau\right)}{\theta\left(\pm u(z) - d;\tau\right)}.$$

Then on $(-\infty, a_1)$ we have $u^+(z) = u^-(z)$. And on $(\mathrm{b}_{g+1}, \infty)$ we have

$$u^+(z) - u^-(z) = \left(\oint_{\mathcal{C}} \mathrm{d}\omega_j\right)_{j=1}^{g},$$

where $\mathcal{C}$ is a clockwise-oriented simple contour that encircles $[\mathrm{a}_1, \mathrm{b}_{g+1}]$. Then because all the differentials $\mathrm{d}\omega_j$ are of the form $P(z)/R(z)$ where P is a degree $g-1$ polynomial and

$R(z) = \mathrm{O}(z^{g+1})$ as $z \to \infty$, we see that $\oint_{\mathcal{C}} \mathrm{d}\omega_j = 0$. Thus, ignoring any poles Θ may have, we find that Θ satisfies the following jump conditions:

$$\Theta^+(z) = \begin{cases} \Theta^-(z) \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} & z \in (\mathrm{a}_j, \mathrm{b}_j), \\ \Theta^-(z) \begin{bmatrix} \mathrm{e}^{-V_j} & 0 \\ 0 & \mathrm{e}^{V_j} \end{bmatrix} & z \in (\mathrm{b}_j, \mathrm{a}_{j+1}), \\ \Theta^-(z) & z \in (-\infty, \mathrm{a}_1) \cup (\mathrm{b}_{g+1}, \infty). \end{cases}$$

Also, note that since $u(\infty)$ is well-defined, Θ has a limit as $z \to \infty$ and is analytic at infinity.

Of particular importance are the poles of Θ. It is known that (see [7], for example) if $\theta(u(P) - \mathcal{A}(D) - \mathbf{k}), D = P_1 + \cdots + P_g$, is not identically zero (This holds if D is nonspecial.) , then, counting multiplicities, $\theta(u(P) - \mathcal{A}(D) - \mathbf{k})$, has g zeros on Σ. These zeros are characterized by

$$\theta(u(P) - \mathcal{A}(D) - \mathbf{k}) = 0 \quad \Leftrightarrow \quad P = P_j,$$

for some j. Next, define

$$\gamma(z) = \left[\prod_{j=1}^{g+1} \left(\frac{z - \mathrm{b}_j}{z - \mathrm{a}_j}\right)\right]^{1/4}, \tag{A.3}$$

analytic on $\mathbb{C} \setminus \cup_j [\mathrm{a}_j, \mathrm{b}_j]$, with $\gamma(z) \sim 1, z \to \infty$. It follows that $\gamma - \gamma^{-1}$ has a single root z_j in $(\mathrm{b}_j, \mathrm{a}_{j+1})$ for $j = 1, 2, \ldots, g$, while $\gamma + \gamma^{-1}$ does not vanish on $\mathbb{C} \setminus \cup_j [\mathrm{a}_j, \mathrm{b}_j]$. So, define two divisors

$$D_1 = \sum_{j=1}^{g} \pi_1^{-1}(z_j), \quad D_2 = \sum_{j=1}^{g} \pi_2^{-1}(z_j).$$

It follows from [35] (see also [73, Lemma 11.10]) that these divisors are nonspecial and therefore the θ functions we will consider do not vanish identically.

Note that for $\mathbf{d}_1 := \mathcal{A}(D_1) + \mathbf{k}$, the function $z \mapsto \theta(u(z) - \mathbf{d}_1; \tau)$ has zeros at z_j, while the function $z \mapsto \theta(-u(z) - \mathbf{d}_1; \tau)$ is non-vanishing. Similarly, for

$$\mathbf{d}_2 := \mathcal{A}(D_2) + \mathbf{k}, \tag{A.4}$$

the function $z \mapsto \theta(-u(z) - \mathbf{d}_2; \tau)$ has zeros at z_j, while the function $z \mapsto \theta(u(z) - \mathbf{d}_2; \tau)$ is non-vanishing.

Inspired by [30], this leads us to consider

$$L_n(z) = \begin{bmatrix} \left(\frac{\gamma(z)+\gamma(z)^{-1}}{2}\right)\Theta_1(z;\mathbf{d}_2;v) & \left(\frac{\gamma(z)-\gamma(z)^{-1}}{2\mathrm{i}}\right)\Theta_2(z;\mathbf{d}_2;v) \\ \left(\frac{\gamma(z)^{-1}-\gamma(z)}{2\mathrm{i}}\right)\Theta_1(z;\mathbf{d}_1;v) & \left(\frac{\gamma(z)+\gamma(z)^{-1}}{2}\right)\Theta_2(z;\mathbf{d}_1;v) \end{bmatrix}, \tag{A.5}$$

which is analytic in $\mathbb{C}\setminus\cup_j[\mathrm{a}_j,\mathrm{b}_j]$, with a limit as $z\to\infty$ and satisfies the jumps:

$$L_n^+(z) = \begin{cases} L_n^-(z)\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} & z\in(\mathrm{a}_j,\mathrm{b}_j), \\ L_n^-(z)\begin{bmatrix} e^{-v_j} & 0 \\ 0 & e^{v_j} \end{bmatrix} & z\in(\mathrm{b}_j,\mathrm{a}_{j+1}), \\ L_n^-(z) & z\in(-\infty,\mathrm{a}_1)\cup(\mathrm{b}_{g+1},\infty). \end{cases}$$

This follows because $\gamma^+(z) = \mathrm{i}\gamma^-(z)$ for $z\in(\mathrm{a}_j,\mathrm{b}_j)$ and therefore

$$\gamma^+(z) + (\gamma(z)^{-1})^+ = \mathrm{i}\big(\gamma^-(z) - (\gamma(z)^{-1})^-\big),$$
$$\gamma^+(z) - (\gamma(z)^{-1})^+ = \mathrm{i}\big(\gamma^-(z) + (\gamma(z)^{-1})^-\big).$$

We point out that (A.5) was first proposed in [30] and then used by many authors, to list but a few, [17, 20, 21].

A.2 Asymptotics of orthogonal polynomials: proof of Theorem 2.2

The derivation of the asymptotic formulae proceeds in six steps, each of which transforms $Y_n(z;\mu)$ by explicit algebraic transformations:

- Step 1: Turn residue conditions into rational jump conditions.
- Step 2: The determination of a differential, also called the exterior Green's function with pole at infinity, that is used to remove the singularities of Y_n at infinity.
- Step 3: Lens the Riemann–Hilbert problem, invoking analyticity of functions in the jump matrix, to judiciously factor and move jumps into regions where exponential decay can be induced.
- Step 4: Use the differential to remove the singularities at infinity and induce exponential decay (decay to the identity matrix) on contours moved away the support of μ.

- Step 5: Determine the Szegoő function that removes the details of the remaining jumps and converts them to piecewise constant jumps.
- Step 6: Now that the original unknown Y_n has been transformed to something that has jump matrices that are exponentially close to being piecewise constant, the limiting "model" Riemann–Hilbert problem is solved explicitly using theta functions.

The result, after unwinding all the transformations, is an explicit asymptotic expression for Y_n with exponentially small error terms. This procedure is far from new as it is applied in this form to measures supported on a single interval in [1, 55, 57] and in greater generality in [77]. We rederive the results of [77] in our special case to make them more explicit.

A.2.1 *Step 1: residue conditions to rational jumps*

Consider the function $Y_n(z;\mu)$ as defined in (2.2). Now, consider a new unknown,

$$Z_n(z;\mu) = Y_n(z;\mu)\begin{bmatrix} \prod_{j=1}^{p}(z-\mathrm{c}_j)^{-1} & 0 \\ 0 & \prod_{j=1}^{p}(z-\mathrm{c}_j) \end{bmatrix}.$$

This eliminates poles in the second column and adds them to the first. The residue condition implies that near c_j

$$Y_n(z;\mu) = \begin{bmatrix} Y_n(\mathrm{c}_j;\mu)_{11} + \mathrm{O}(z-\mathrm{c}_j) & \frac{w_j}{2\pi\mathrm{i}}\frac{Y_n(\mathrm{c}_j;\mu)_{11}}{z-\mathrm{c}_j} + \mathrm{O}(1) \\ Y_n(\mathrm{c}_j;\mu)_{21} + \mathrm{O}(z-\mathrm{c}_j) & \frac{w_j}{2\pi\mathrm{i}}\frac{Y_n(\mathrm{c}_j;n)_{21}}{z-\mathrm{c}_j} + \mathrm{O}(1) \end{bmatrix}.$$

Then, for Z_n, we have

$$Z_n(z;\mu) = \begin{bmatrix} \frac{Y_n(\mathrm{c}_j;\mu)_{11}}{z-\mathrm{c}_j}\prod_{k\neq j}(\mathrm{c}_j-\mathrm{c}_k)^{-1} + \mathrm{O}(1) & \frac{w_j}{2\pi\mathrm{i}}Y_n(\mathrm{c}_j;\mu)_{11}\prod_{k\neq j}(\mathrm{c}_j-\mathrm{c}_k) + \mathrm{O}(z-\mathrm{c}_j) \\ \frac{Y_n(\mathrm{c}_j;\mu)_{21}}{z-\mathrm{c}_j}\prod_{k\neq j}(\mathrm{c}_j-\mathrm{c}_k)^{-1} + \mathrm{O}(1) & \frac{w_j}{2\pi\mathrm{i}}Y_{21}(\mathrm{c}_j;n)\prod_{k\neq j}(\mathrm{c}_j-\mathrm{c}_k) + \mathrm{O}(z-\mathrm{c}_j) \end{bmatrix}.$$

From this, it follows that

$$\mathrm{Res}_{z=\mathrm{c}_j} Z_n(z;\mu) = \begin{bmatrix} Y_n(\mathrm{c}_j;\mu)_{11}\prod_{k\neq j}(\mathrm{c}_j-\mathrm{c}_k)^{-1} & 0 \\ Y_n(\mathrm{c}_j;\mu)_{21}\prod_{k\neq j}(\mathrm{c}_j-\mathrm{c}_k)^{-1} & 0 \end{bmatrix} = \lim_{z\to\mathrm{c}_j} Z_n(z;\mu)\begin{bmatrix} 0 & 0 \\ \frac{2\pi\mathrm{i}}{w_j}\prod_{k\neq j}(\mathrm{c}_j-\mathrm{c}_k)^{-2} & 0 \end{bmatrix}.$$

The other important properties of $Z_n(z;\mu)$ are given by

$$\lim_{\epsilon\to 0^+} Z_n(z+\mathrm{i}\epsilon;\mu) = \lim_{\epsilon\to 0^+} Z_n(z-\mathrm{i}\epsilon;\mu)\begin{bmatrix} 1 & \rho(z)\prod_{j=1}^{p}(z-\mathrm{c}_j)^2 \\ 0 & 1 \end{bmatrix}, \quad z\in(-1,1),$$

$$Z_n(z;\mu)\begin{bmatrix} z^{-(n-p)} & 0 \\ 0 & z^{n-p} \end{bmatrix} = I + \mathrm{O}(1/z), \quad z\to\infty.$$

Now, let Σ_j be a small circle centered at c_j with radius sufficiently small so that it does not intersect any other Σ_k for $k \neq j$ and so that it does not intersect any Σ_j for all j. Denote by $\mathring{\Sigma}_j$ the region enclosed by Σ_j. Define

$$\check{Z}_n(z;\mu) = \begin{cases} Z_n(z;\mu) & z \in \mathbb{C} \setminus \left(\bigcup_{j=1}^{g+1} [\mathrm{a}_j, \mathrm{b}_j] \cup \bigcup_{j=1}^{p} (\Sigma_j \cup \mathring{\Sigma}_j) \right), \\ Z(z;n) \begin{bmatrix} 1 & 0 \\ -\frac{\tilde{w}_j}{z-\mathrm{c}_j} & 1 \end{bmatrix} & z \in \mathring{\Sigma}_j \setminus \{\mathrm{c}_j\}, \end{cases}$$

where $\tilde{w}_j$ is defined as

$$\tilde{w}_j := \frac{2\pi \mathrm{i}}{w_j} \prod_{k \neq j} (\mathrm{c}_j - \mathrm{c}_k)^{-2}. \tag{A.6}$$

Then it follows that $\check{Z}_n(z;\mu)$ has a removable singularity at $z = \mathrm{c}_j$ for each j. We give Σ_j counter-clockwise orientation and denote by $\check{Z}_n^{\pm}$ the limit to Σ_j from the interior $(+)$ or exterior $(-)$. We have

$$\check{Z}_n^{+}(z;\mu) = \check{Z}_n^{-}(z;\mu) \begin{bmatrix} 1 & 0 \\ \frac{\tilde{w}_j}{z-\mathrm{c}_j} & 1 \end{bmatrix}, \quad z \in \Sigma_j.$$

A.2.2 *Step 2: determine the correct differential*

Our next task is to remove the growth/decay at infinity. We look for a function $\mathfrak{g}$ that satisfies:

(a) $\mathfrak{g}'(z) = 1/z + O(1/z^2)$ as $z \to \infty$.

(b) $\mathfrak{g}'_+(z), \mathfrak{g}'_-(z) \in \mathrm{i}\mathbb{R}$ on $[\mathrm{a}_j, \mathrm{b}_j]$.

(c) $\int_{\mathrm{b}_j}^{\mathrm{a}_{j+1}} \mathfrak{g}'(z) \mathrm{d}z = 0, j = 1, 2, \ldots, g$

Based on this, define

$$\mathfrak{g}'(z) = \frac{Q_g(z)}{R(z)}, \quad \text{where} \quad R(z)^2 = \prod_{j=1}^{g+1} (z - \mathrm{a}_j)(z - \mathrm{b}_j), \tag{A.7}$$

where Q_g is a monic polynomial of degree g, providing g degrees of freedom to satisfy the requisite conditions. We then see that $R_+(z)$ is purely imaginary in each interval $(\mathrm{a}_j, \mathrm{b}_j)$ and real-valued on $(\mathrm{b}_j, \mathrm{a}_{j+1})$. The linear system that defines $Q_g(z) = \sum_k h_k z^k$ is given by:

$$\int_{\mathrm{b}_j}^{\mathrm{a}_{j+1}} \sum_{k=0}^{g-1} h_k \frac{z^k}{R(z)} \mathrm{d}z = -\int_{\mathrm{b}_j}^{\mathrm{a}_{j+1}} \frac{z^g}{R(z)} \mathrm{d}z, \quad j = 1, 2, \ldots, g.$$

Therefore, h_k are real-valued coefficients. This implies (b). The unique solvability of this system for these coefficients follows from the fact that $\frac{z^k}{R(z)} dz$, $k = 0, 1, 2, \ldots, g-1$ forms a basis for holomorphic differentials on hyperelliptic Riemann surface defined by

$w^2 = R(z)^2$. Then because $R(z)$ is sign definite in each gap $(\mathrm{b}_j, \mathrm{a}_{j+1})$, for (c) to hold, $\mathfrak{g}'(z)$ must vanish in this interval. This implies that $Q_g(z)$ has one root d_j in each gap $(\mathrm{b}_j, \mathrm{a}_{j+1})$ and this accounts for all the roots of $Q_g(z)$. This implies that $\mathfrak{g}'(z) < 0$ for $z < \mathrm{a}_1$ and $\mathfrak{g}'(z) > 0$ for $z > \mathrm{b}_{g+1}$. With the notation $\mathrm{b}_0 = -\infty$ and $\mathrm{a}_{g+2} = +\infty$, it follows that

$$R(z)R(z') < 0, \quad z \in (\mathrm{b}_j, \mathrm{a}_{j+1}), \quad z' \in (\mathrm{b}_{j+1}, \mathrm{a}_{j+2}),$$

for $j = 0, 1, 2, \ldots, g-1$. Since $\mathfrak{g}'(z) < 0$ for $z < \mathrm{a}_1$, we see that $\mathfrak{g}'(z) > 0$ for $z \in (\mathrm{b}_1, \mathrm{d}_1)$ and $g'(z) < 0$ for $z \in (\mathrm{d}_1, \mathrm{a}_2)$. This is true, in general, with $\mathfrak{g}'(z)$ being positive on $(\mathrm{b}_j, \mathrm{d}_j)$ and negative on $(\mathrm{d}_j, \mathrm{a}_{j+1})$.

Then $\mathfrak{g}(z)$ is defined by integration of $\mathfrak{g}'(z)$ from a_1 to z by a straight line. We can compute

$$\mathfrak{g}^+(z) + \mathfrak{g}^-(z) = 0, \quad z \in (\mathrm{a}_j, \mathrm{b}_j),$$

where we use the fact that $R^+(z) = -R^-(z)$ for $z \in (\mathrm{a}_j, \mathrm{b}_j)$ along with $\int_{\mathrm{b}_j}^{\mathrm{a}_{j+1}} \mathfrak{g}'(z)\mathrm{d}z = 0$ for each j. And for $z \in (\mathrm{b}_j, \mathrm{a}_{j+1})$ we find

$$\mathfrak{g}^+(z) - \mathfrak{g}^-(z) = 2\sum_{k=1}^{j} \int_{\mathrm{a}_k}^{\mathrm{b}_k} (\mathfrak{g}')^+(z)\mathrm{d}z =: \Delta_j. \tag{A.8}$$

So this is constant in each gap $(\mathrm{b}_j, \mathrm{a}_{j+1})$ and is purely imaginary. Define the vector

$$\boldsymbol{\Delta} = (\Delta_j)_{j=1}^{g}. \tag{A.9}$$

It is easy to see that the above arguments result in the following proposition.

Proposition A.1. For $\mathfrak{g}(z)$ defined in (A.7), we have that $\operatorname{Re}\mathfrak{g}(z)$ is strictly positive on any closed subset of $\mathbb{R} \setminus \cup_j[\mathrm{a}_j, \mathrm{b}_j]$.

Combining Proposition A.1 with the maximum modulus principle applied to $\mathrm{e}^{-\mathfrak{g}(z)}$, this statement extends to $\mathbb{C} \setminus \cup_j[\mathrm{a}_j, \mathrm{b}_j]$. Define

$$\mathfrak{c} = \lim_{z\to\infty} \frac{\mathrm{e}^{\mathfrak{g}(z)}}{z}. \tag{A.10}$$

We remark that $|\mathfrak{c}|$ is classically known as the capacity of $\cup_j[\mathrm{a}_j, \mathrm{b}_j]$ [65].

A.2.3 *Step 3: lens the problem*

Define $\check{\rho}_j$ to be the analytic continuation of $\rho(z)\prod_{j=1}^{p}(z - \mathrm{c}_j)^2$ off $[\mathrm{a}_j, \mathrm{b}_j]$ to Ω_j. Then let C_j be a curve the encircles $[\mathrm{a}_j, \mathrm{b}_j]$ lying in Ω_j. Denote the interior of this curve by D_j.

Then define

$$S_n(z;\mu) = \begin{cases} \check{Z}_n(z;\mu)\begin{bmatrix} 1 & 0 \\ -1/\check{\rho}_j(z) & 1 \end{bmatrix} & z \in D_j \cap \mathbb{C}^+, \\ \check{Z}_n(z;\mu)\begin{bmatrix} 1 & 0 \\ 1/\check{\rho}_j(z) & 1 \end{bmatrix} & z \in D_j \cap \mathbb{C}^-, \\ \check{Z}_n(z;\mu) & \text{otherwise.} \end{cases}$$

We find

$$S_n^+(z;\mu) = \begin{cases} S_n^-(z;\mu)\begin{bmatrix} 1 & 0 \\ 1/\check{\rho}_j(z) & 1 \end{bmatrix} & z \in C_j \setminus \mathbb{R}, \\ S_n^-(z;\mu)\begin{bmatrix} 0 & \check{\rho}_j(z) \\ -1/\check{\rho}_j(z) & 0 \end{bmatrix} & z \in (\mathtt{a}_j, \mathtt{b}_j), \\ S_n^-(z;\mu)\begin{bmatrix} 1 & 0 \\ \frac{\tilde{w}_j}{z-\mathtt{c}_j} & 1 \end{bmatrix} & z \in \Sigma_j \end{cases}$$

Note that S_n still has the same normalization at infinity as $\check{Z}_n$. And recalling that $\check{Z}_n(z;\mu)$ is bounded on D_j, we see that we have now introduced unbounded behavior in S_n, in an entrywise sense,

$$S_n(z;\mu) = \begin{bmatrix} O(|z-\mathtt{a}_j|^{-1/2}) & O(1) \\ O(|z-\mathtt{a}_j|^{-1/2}) & O(1) \end{bmatrix}, \quad S_n(z;\mu) = \begin{bmatrix} O(|z-\mathtt{b}_j|^{-1/2}) & O(1) \\ O(|z-\mathtt{b}_j|^{-1/2}) & O(1) \end{bmatrix},$$

as $z \to \mathtt{a}_j, \mathtt{b}_j$, respectively.

A.2.4 *Step 4: normalize at infinity*

Define

$$\check{S}_n(z;\mu) = \mathfrak{c}^{(n-p)\sigma_3} S_n(z;\mu)\, \mathrm{e}^{-(n-p)\mathfrak{g}(z)\sigma_3}$$

Then it follows that $\check{S}_n(z;\mu) = I + \mathrm{O}(z^{-1})$ as $z \to \infty$ and it satisfies the jumps

$$\check{S}_n^+(z;\mu) = \begin{cases} \check{S}_n^-(z;\mu) \begin{bmatrix} 1 & 0 \\ \mathrm{e}^{-2(n-p)\mathfrak{g}(z)}/\check{\rho}_j(z) & 1 \end{bmatrix} & z \in C_j \setminus \mathbb{R}, \\ \check{S}_n^-(z;\mu) \begin{bmatrix} 0 & \check{\rho}_j(z) \\ -1/\check{\rho}_j(z) & 0 \end{bmatrix} & z \in (\mathrm{a}_j, \mathrm{b}_j), \\ \check{S}_n^-(z;\mu) \begin{bmatrix} \mathrm{e}^{-(n-p)\Delta_j} & 0 \\ 0 & \mathrm{e}^{(n-p)\Delta_j} \end{bmatrix} & z \in (\mathrm{b}_j, \mathrm{a}_{j+1}), \\ \check{S}_n^-(z;\mu) \begin{bmatrix} 1 & 0 \\ \mathrm{e}^{-2(n-p)\mathfrak{g}(z)} \frac{\tilde{w}_j}{z-\mathrm{c}_j} & 1 \end{bmatrix} & z \in \Sigma_j. \end{cases}$$

A.2.5 *Step 5: determine the Szegoő function*

The point of the Szegoő function is to replace the jumps on $(\mathrm{a}_j, \mathrm{b}_j)$ with something simpler at the cost of adding to the jumps on $(\mathrm{b}_j, \mathrm{a}_{j+1})$. Define

$$\mathsf{G}(z) = -\frac{R(z)}{2\pi \mathrm{i}} \left[\sum_{j=1}^{g+1} \int_{\mathrm{a}_j}^{\mathrm{b}_j} \frac{\log \check{\rho}_j(\lambda)}{\lambda - z} \frac{\mathrm{d}\lambda}{R_+(\lambda)} + \sum_{j=1}^{g} \int_{\mathrm{b}_j}^{\mathrm{a}_{j-1}} \frac{\zeta_j}{\lambda - z} \frac{\mathrm{d}\lambda}{R(\lambda)} \right], \tag{A.11}$$

where the constants ζ_j are yet to be determined.

Before we determine these constants, note that

$$\mathsf{G}^+(z) + \mathsf{G}^-(z) = -\log \check{\rho}_j(z), \quad z \in (\mathrm{a}_j, \mathrm{b}_j),$$
$$\mathsf{G}^+(z) - \mathsf{G}^-(z) = -\zeta_j, \quad z \in (\mathrm{b}_j, \mathrm{a}_{j+1}).$$

Since $R(z) = \mathrm{O}(z^g)$, we see that $\mathsf{G}(z) = \mathrm{O}(z^{g-1})$. To avoid unbounded behavior of G at infinity, we choose $\boldsymbol{\zeta} = (\zeta_j)_{j=1}^g$ so that as $z \to \infty$

$$\mathsf{G}(z) = \mathrm{O}(1). \tag{A.12}$$

Indeed, we find a linear system of equations

$$m_\ell = -\sum_{j=1}^{g}\int_{\mathrm{a}_j}^{\mathrm{b}_j} \log\check{\rho}_j(\lambda)\lambda^{\ell-1}\frac{\mathrm{d}\lambda}{R_+(\lambda)} \tag{A.13}$$

$$-\sum_{j=1}^{g-1}\int_{b_j}^{a_{j-1}} \zeta_j\lambda^{\ell-1}\frac{\mathrm{d}\lambda}{R_+(\lambda)} = 0, \quad \ell = 1,2,\ldots,g-1.$$

We pause briefly to discuss the singularity behavior of G and note that we have to take some care because in Assumption 1 we allow μ to depend on N.

Lemma A.2. Given Assumption 1, for some $\epsilon > 0$, and for every $j = 1,2,\ldots,g+1$, we have

$$\mathsf{G}(z) = -\frac{1}{4}\log[(z-\mathrm{b}_j)(\mathrm{a}_j - z)] + \mathtt{R}_j(z), \quad \mathrm{dist}(z,[\mathrm{a}_j,\mathrm{b}_j]) \le \epsilon,$$

where $\mathtt{R}_j(z)$ is a uniformly bounded function for $\mathrm{dist}(z,[\mathrm{a}_j,\mathrm{b}_j]) \le \epsilon$.

Proof. Recall Proposition A.1. We first observe that if h is a uniformly bounded analytic function in the $O_\epsilon = \{z : \mathrm{dist}(z,[\mathrm{a}_j,\mathrm{b}_j]) \le \epsilon\}$ then

$$E(\lambda) = \frac{R(z)}{2\pi\mathrm{i}}\int_{\mathrm{a}_j}^{\mathrm{b}_j} \frac{h(\lambda)}{\lambda - z}\frac{\mathrm{d}\lambda}{R_+(\lambda)} \tag{A.14}$$

is bounded for z in any fixed bounded set. Indeed, for $z \in O_{\epsilon/2}$

$$\frac{R(z)}{2\pi\mathrm{i}}\int_{\mathrm{a}_j}^{\mathrm{b}_j} \frac{h(\lambda)}{\lambda - z}\frac{\mathrm{d}\lambda}{R_+(\lambda)} = \frac{h(z)}{2} - \frac{R(z)}{4\pi\mathrm{i}}\int_\Sigma \frac{h(z')}{z'-z}\frac{\mathrm{d}z'}{R(z')},$$

where $\Sigma = \partial O_{2\epsilon/3}$. This is uniformly bounded. This function is then evidently bounded uniformly on $\{|z| \le R\} \setminus O_{\epsilon/2}$ for any $R > 0$. So, consider

$$H(z) = \frac{R(z)}{2\pi\mathrm{i}}\int_{\mathrm{a}_j}^{\mathrm{b}_j} \frac{\frac{1}{2}\log(\lambda - \mathrm{a}_j)_+}{\lambda - z}\frac{\mathrm{d}\lambda}{R_+(\lambda)},$$

in a neighborhood of a_j where the branch cut of $\log z$ here is chosen to be $[0,\infty)$. By choosing ϵ sufficiently small, we find

$$H(z) = \frac{1}{4}\log(z-\mathrm{a}_j) - \frac{R(z)}{4\pi\mathrm{i}}\int_{\partial\tilde{O}_{2\epsilon}} \frac{\frac{1}{2}\log(z'-\mathrm{a}_j)}{z'-z}\frac{\mathrm{d}z'}{R(z)} + E(z),$$

where $\tilde{O}_{2\epsilon} = O_{2\epsilon} \cap \{\operatorname{Re} z \le \mathrm{b}_j\}$ and $h = -\pi\mathrm{i}/2$ in (A.14) is a constant. This holds for $z \in O_\epsilon \cap \{\operatorname{Re} z \le \mathrm{b}_j - \epsilon\}$. The second two terms are uniformly bounded for these values of z. We exchange $\log z$ for the principal branch in the initial integral for $H(z)$ and find

$$H(z) = \frac{1}{4}\log(z - \mathrm{a}_j) - \frac{R(z)}{4\pi\mathrm{i}}\int_{\partial\check{O}_{2\epsilon}} \frac{\frac{1}{2}\log(z' - \mathrm{a}_j)}{z' - z}\frac{\mathrm{d}z'}{R(z)} - E(z),$$

where $\check{O}_{2\epsilon} = O_{2\epsilon} \cap \{\operatorname{Re} z \ge \mathrm{a}_j\}$. The second two terms are uniformly bounded for $z \in O_\epsilon \cap \{\operatorname{Re} z \ge \mathrm{a}_j + \epsilon\}$. Similar arguments hold after exchanging $\log(\lambda - \mathrm{a}_j)$ for $\log(\lambda - \mathrm{b}_j)$, and the facts about $E(\lambda)$ apply to the integral of each of ζ_j, and the lemma follows. ∎

This system of equations is uniquely solvable for $\boldsymbol{\zeta}$ using the fact that the normalized differentials exist, and involves the same coefficient matrix that is used to determine the polynomials Q_g above.

Then consider

$$T_n(z;\mu) = \mathrm{e}^{\sigma_3 \mathsf{G}(\infty)}\,\check{S}_n(z;\mu)\,\mathrm{e}^{-\sigma_3 \mathsf{G}(z)}.$$

We check the jumps of T_n:

$$T_n^+(z;\mu) = \begin{cases} T_n^-(z;\mu)\begin{bmatrix} 1 & 0 \\ \mathrm{e}^{-2((n-p)\mathfrak{g}(z) - \mathsf{G}(z))}/\check{\rho}_j(z) & 1 \end{bmatrix} & z \in C_j \setminus \mathbb{R}, \\ T_n^-(z;\mu)\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} & z \in (\mathrm{a}_j, \mathrm{b}_j), \\ T_n^-(z;\mu))\begin{bmatrix} \mathrm{e}^{-(n-p)\Delta_j - \zeta_j} & 0 \\ 0 & \mathrm{e}^{(n-p)\Delta_j + \zeta_j} \end{bmatrix} & z \in (\mathrm{b}_j, \mathrm{a}_{j+1}), \\ T_n^-(z;\mu)\begin{bmatrix} 1 & 0 \\ \mathrm{e}^{-2((n-p)\mathfrak{g}(z) - \mathsf{G}(z))}\frac{\tilde{w}_j}{z - \mathrm{c}_j} & 1 \end{bmatrix} & z \in \Sigma_j. \end{cases}$$

Since the first and last jumps tend to the identity matrix exponentially fast, uniformly at a rate $O(e^{-cn})$ for some $c > 0$ and we look to solve the model problem

$$\check{T}_n^+(z;\mu) = \begin{cases} \check{T}_n^-(z;\mu) \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} & z \in (\mathsf{a}_j, \mathsf{b}_j), \\ \check{T}_n^-(z;\mu) \begin{bmatrix} e^{-(n-p)\Delta_j - \zeta_j} & 0 \\ 0 & e^{(n-p)\Delta_j + \zeta_j} \end{bmatrix} & z \in (\mathsf{b}_j, \mathsf{a}_{j+1}), \end{cases}$$

with the condition that $\check{T}_n(\infty;\mu) = I$.

A.2.6 *Step 6: solution of the model problem*

From (A.5), we find that $\check{T}_n(z;\mu) = L_n(\infty)^{-1} L_n(z)$, with $v_j = (n-p)\Delta_j + \zeta_j, j = 1, 2, \dots, g$, that is, $v = (n-p)\boldsymbol{\Delta} + \boldsymbol{\zeta}$. It then follows that

$$R_n(z;\mu) := T_n(z;\mu)\check{T}_n(z;\mu)^{-1},$$

using the fact that $\check{T}_n(z;\mu)$ and its inverse are uniformly bounded (see [21], for example) on sets bounded away from the support of μ, it follows that

$$R_n(z;\mu) = I + O\left(\frac{e^{-cn}}{1+|z|}\right).$$

Appendix B. Algorithmic Asymptotic Expansions: Proof of Theorem 3.1 and Its Corollaries

B.1 Detailed expressions of (2.17)

We provide some explicit entry-wise formulae for (2.17). Denote

$$E(z;n) = \left(I + O\left(\frac{e^{-cn}}{1+|z|}\right)\right) L_n(\infty)^{-1} L_n(z). \tag{B.1}$$

According to (2.17), we readily obtain that for z outside any region of deformation

$$Y_n(z;\mu)_{11} = \mathfrak{c}^{(p-n)}\, e^{-\mathsf{G}(\infty)}\, e^{\mathsf{G}(z)}\, e^{(n-p)\mathfrak{g}(z)} \left[\prod_{j=1}^{p} (z - \mathsf{c}_j)\right] E_{11}(z;n), \tag{B.2}$$

$$Y_n(z;\mu)_{12} = \mathfrak{c}^{(p-n)}\, e^{-\mathsf{G}(\infty)}\, e^{-\mathsf{G}(z)}\, e^{-(n-p)\mathfrak{g}(z)} \left[\prod_{j=1}^{p} (z - \mathsf{c}_j)^{-1}\right] E_{12}(z;n), \tag{B.3}$$

$$Y_n(z;\mu)_{21} = \mathfrak{c}^{-(p-n)}\, \mathrm{e}^{\mathsf{G}(\infty)}\, \mathrm{e}^{\mathsf{G}(z)}\, \mathrm{e}^{(n-p)\mathfrak{g}(z)} \left[\prod_{j=1}^{p}(z-\mathrm{c}_j)\right] E_{21}(z;n), \tag{B.4}$$

$$Y_n(z;\mu)_{22} = \mathfrak{c}^{-(p-n)}\, \mathrm{e}^{\mathsf{G}(\infty)}\, \mathrm{e}^{-\mathsf{G}(z)}\, \mathrm{e}^{-(n-p)\mathfrak{g}(z)} \left[\prod_{j=1}^{p}(z-\mathrm{c}_j)^{-1}\right] E_{22}(z;n). \tag{B.5}$$

Recall (A.3). As $\gamma(z) \to 1$ when $z \to \infty$, using the definition of $L_n(z)$ in (A.5), we see that

$$L_n(\infty)^{-1}L_n(z) = \begin{bmatrix} \Theta_1(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})^{-1} & 0 \\ 0 & \Theta_2(\infty;\mathbf{d}_1,;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})^{-1} \end{bmatrix}$$
$$\times \begin{bmatrix} \left(\frac{\gamma(z)+\gamma(z)^{-1}}{2}\right)\Theta_1(z;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta}) & \left(\frac{\gamma(z)-\gamma(z)^{-1}}{2\mathrm{i}}\right)\Theta_2(z;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta}) \\ \left(\frac{\gamma(z)^{-1}-\gamma(z)}{2\mathrm{i}}\right)\Theta_1(z;\mathbf{d}_1;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta}) & \left(\frac{\gamma(z)+\gamma(z)^{-1}}{2}\right)\Theta_2(z;\mathbf{d}_1;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta}) \end{bmatrix}. \tag{B.6}$$

Finally, we provide more explicit formulae for E in (B.1). Note that

$$E_{11}(0;n) = \frac{1}{2}\left(\prod_{j=1}^{g+1}\left(\frac{\mathrm{b}_j}{\mathrm{a}_j}\right)^{1/4} + \prod_{j=1}^{g+1}\left(\frac{\mathrm{a}_j}{\mathrm{b}_j}\right)^{1/4}\right)\frac{\Theta_1(0;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})}{\Theta_1(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})} + \mathrm{O}(\mathrm{e}^{-cn}), \tag{B.7}$$

and similar expressions are easily derivable for the other entries of $E(0;n)$.

Define $\Theta_1^{(1)}$ by,

$$\Theta_1(z;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta}) = \Theta_1(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta}) + \frac{1}{z}\Theta_1^{(1)}(\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta}) + \mathrm{O}(z^{-2}), \tag{B.8}$$

so that $\Theta_1^{(1)}$ denotes the residue of Θ_1 at infinity. Together with (B.6), as $z \to \infty$, it leads to

$$E_{11}(z;n) = 1 + \frac{1}{z}\left[\frac{\Theta_1^{(1)}(\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})}{\Theta_1(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})}\right] + \mathrm{O}(\mathrm{e}^{-cn} + z^{-2}). \tag{B.9}$$

Moreover, using (A.13), we see that as $z \to \infty$

$$\mathrm{e}^{-\mathsf{G}(\infty)}\,\mathrm{e}^{\mathsf{G}(z)} = 1 + \frac{1}{z}\left[\frac{m_{g+1}}{2\pi\mathrm{i}} - \frac{m_g}{2\pi\mathrm{i}}\sum_{j=1}^{g+1}(\mathrm{a}_j+\mathrm{b}_j)\right] + \mathrm{O}(z^{-2}),$$

$$\mathfrak{c}^{(p-n)}\frac{\mathrm{e}^{(n-p)\mathfrak{g}(z)}}{z^n}\left[\prod_{j=1}^{p}(z-\mathrm{c}_j)\right] = 1 + \frac{1}{z}\left[\mathfrak{g}_1 - \sum_{j=1}^{p}\mathrm{c}_j\right] + \mathrm{O}(z^{-2}),$$

where $\mathfrak{g}_1$ is defined so that $\mathfrak{g}(z) = \log z + \log \mathfrak{c} + \mathfrak{g}_1/z + O(z^2)$ as $z \to \infty$. Combining (B.13),(B.8), (B.2), (B.6), (A.10), and

$$z\left[z^{-n}Y_n(z;\mu)_{11} - 1\right] = z\left[\mathfrak{c}^{(p-n)}\,\mathrm{e}^{-\mathsf{G}(\infty)}\,\mathrm{e}^{\mathsf{G}(z)}\,\frac{\mathrm{e}^{(n-p)\mathfrak{g}(z)}}{z^n}\left[\prod_{j=1}^{p}(z-\mathrm{c}_j)\right]E_{11}(z;n) - 1\right],$$

one finds

$$\lim_{z\to\infty} z\left(z^{-n}Y_n(z;\mu)_{11} - 1\right) = \frac{m_{g+1}}{2\pi\mathrm{i}} - \frac{m_g}{2\pi\mathrm{i}}\sum_{j=1}^{g+1}(\mathrm{a}_j+\mathrm{b}_j) + (n-p)\mathfrak{g}_1 - \sum_{j=1}^{p}\mathrm{c}_j$$
$$+\frac{\Theta_1^{(1)}(\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})}{\Theta_1(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})} + O(\mathrm{e}^{-cn}).$$

Also, according to (B1) and (B.6), we see that

$$\lim_{z\to\infty} zE_{12}(z;n) = \frac{\mathrm{i}}{4}\sum_{j=1}^{g+1}(\mathrm{b}_j-\mathrm{a}_j)\frac{\Theta_2(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})}{\Theta_1(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})} + O(\mathrm{e}^{-cn}), \tag{B.10}$$

where we used the definition (A.3). Consequently, using (B.3), we readily obtain that

$$-2\pi\mathrm{i}\lim_{z\to\infty} z^{n+1}Y_n(z;\mu)_{12} = \left[\frac{\pi}{2}\sum_{j=1}^{g+1}(\mathrm{b}_j-\mathrm{a}_j)\frac{\Theta_2(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})}{\Theta_1(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})} + O(\mathrm{e}^{-cn})\right]$$
$$\times \lim_{z\to\infty} e^{-\mathsf{G}(\infty)-\mathsf{G}(z)}\mathfrak{c}^{p-n}\frac{z^{n-p}}{e^{(n-p)\mathfrak{g}(z)}}\frac{z^p}{\prod_{j=1}^{p}(z-\mathrm{c}_j)^{-1}}$$
$$= e^{-2\mathsf{G}(\infty)}\mathfrak{c}^{2(p-n)}\frac{\pi}{2}\sum_{j=1}^{g+1}(\mathrm{b}_j-\mathrm{a}_j)\frac{\Theta_2(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})}{\Theta_1(\infty;\mathbf{d}_2;(n-p)\boldsymbol{\Delta}+\boldsymbol{\zeta})} + O(\mathrm{e}^{-cn}),$$

where in the last step we used the definition (A.10).

B.2 Asymptotic formulae of Section 3.2

Proof of Corollary 3.2. Recall (3.3). By equating coefficients in (3.4) and the definition of $p_n(z;\mu)$, we find that

$$\ell_n(\mu) = b_n(\mu)\ell_{n+1}(\mu), \tag{B.11}$$
$$s_n(\mu) = a_n(\mu)\ell_n(\mu) + b_n(\mu)s_{n+1}(\mu).$$

There is then, of course, the relation $\gamma_n(\mu) = -2\pi i \ell_n^2(\mu)$. A direct calculation, using orthogonality and the definitions (2.3) and (2.1), leads to

$$\lim_{z\to\infty} z^{n+1} c_n(z;\mu) = -\frac{1}{2\pi i} \lim_{z\to\infty} z^n \int \frac{\pi_n(x;\mu)}{1-(x/z)} dx = -\frac{1}{2\pi i} \int x^n \pi_n(x;\mu) dx$$
$$= -\frac{1}{2\pi i} \ell_n^{-2}(\mu).$$

This gives

$$b_n(\mu)^2 = \frac{\gamma_n(\mu)}{\gamma_{n+1}(\mu)} = \frac{\lim_{z\to\infty} z^{n+2} c_{n+1}(z;\mu)}{\lim_{z\to\infty} z^{n+1} c_n(z;\mu)} = \lim_{z\to\infty} \frac{zY_{n+1}(z;\mu)_{12}}{Y_n(z;\mu)_{12}}. \tag{B.12}$$

The expression

$$a_n(\mu) = \lim_{z\to\infty} z^{-n+1}(\pi_n(z;\mu) - z^n) - \lim_{z\to\infty} z^{-n}(\pi_{n+1}(z;\mu) - z^{n+1}),$$

directly follows from (B.11) and the definition of s_n, ℓ_n. From the definition (2.17), one has

$$a_n(\mu) = -\lim_{z\to\infty} z^{-n}(Y_{n+1}(z;\mu)_{11} - zY_n(z;\mu)_{11}). \tag{B.13}$$

Then the proof follows immediately from Theorem 3.1. ■

Proof of Corollary 3.3. By [64, Proposition 4.1] and the definition (2.17), we have

$$\|\mathbf{e}_n\|_W^2 = 2\pi i \frac{c_n(0;\mu)}{\pi_n(0;\mu)} = 2\pi i \frac{Y_n(0;\mu)_{12}}{Y_n(0;\mu)_{11}}, \tag{B.14}$$

and

$$\|\mathbf{r}_n\|_2^2 = \frac{\prod_{j=0}^{n-1} b_j(\mu)^2}{\pi_n(0;\mu)^2} = \frac{-2\pi i \lim_{z\to\infty} z^{n+1} Y_n(z;\mu)_{12}}{Y_n(0;\mu)_{11}^2}, \tag{B.15}$$

where we used (B.12):

$$\prod_{j=0}^{n-1} b_j(\mu)^2 = \lim_{z\to\infty} \prod_{j=0}^{n-1} \frac{zY_{j+1}(z;\mu)_{12}}{Y_j(z;\mu)_{12}} = \lim_{z\to\infty} z^n \frac{Y_n(z;\mu)_{12}}{Y_0(z;\mu)_{12}},$$

and that $Y_0(z;\mu)_{12} = -\frac{1}{2\pi i z}(1 + O(z^{-1}))$.

The proof of the first equation follows directly from the above formula and (B.2) and (B.3). For the second equation, combining with Theorem 3.1, we can complete the proof. ■

Proof of Corollary 3.4. Using the facts

$$\det \mathcal{J}_n(\mu) = \prod_{j=0}^{n-1} \alpha_j(\mu)^2, \quad \pi_n(z;\mu) = \det(zI - \mathcal{J}_n(\mu)),$$

we obtain that

$$(-1)^n \pi_n(0;\mu) = \prod_{j=0}^{n-1} \alpha_j(\mu)^2.$$

Combining with $(\mathcal{J})_{j,j+1} = \alpha_j \beta_j$, we immediately see that

$$\prod_{j=0}^{n-1} \frac{\beta_j(\mu)^2}{\alpha_j(\mu)^2} = \frac{\prod_{j=0}^{n-1} b_j(\mu)^2}{\pi_n(0;\mu)^2}.$$

This gives the expressions

$$\alpha_n(\mu)^2 = -\frac{\pi_{n+1}(0;\mu)}{\pi_n(0;\mu)}, \quad \beta_n(\mu)^2 = -b_n^2(\mu)\frac{\pi_n(0;\mu)}{\pi_{n+1}(0;\mu)}. \tag{B.16}$$

The proof then follows from Theorem 3.1 and Corollary 3.2. ■

Appendix C. CLT for Spiked Sample Covariance Matrix Model: Proof of Theorem 4.3

In this section, we prove the CLT as in Section 4.4. Throughout this section, we will consistently use the notion of *stochastic domination*, which provides a precise statement of the form "ξ_N is bounded by ζ_N up to a small power of N with high probability".

Definition 2. (i) Let

$$\xi = \left(\xi^{(N)}(u) : N \in \mathbb{N}, u \in U^{(N)}\right), \quad \zeta = \left(\zeta^{(N)}(u) : N \in \mathbb{N}, u \in U^{(N)}\right)$$

be two families of nonnegative random variables, where $U^{(N)}$ is a possibly n-dependent parameter set. We say ξ is stochastically dominated by ζ, uniformly in u, if for any fixed (small) $\epsilon > 0$ and (large) $D > 0$,

$$\sup_{u \in U^{(N)}} \mathbb{P}\left(\xi^{(N)}(u) > N^{\epsilon}\zeta^{(N)}(u)\right) \leq N^{-D},$$

for large enough $N \geq N_0(\epsilon, D)$, and we shall use the notation $\xi \prec \zeta$. Throughout this paper, the stochastic domination will always be uniform in all parameters that are not explicitly fixed (such as matrix indices, and z that takes values in some compact set). Note that $N_0(\epsilon, D)$ may depend on quantities that are explicitly constant, such as τ in Assumption 2. If for some complex family ξ we have $|\xi| \prec \zeta$, then we will also write $\xi \prec \zeta$ or $\xi = \mathrm{O}_{\prec}(\zeta)$.

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

C.1 Technical tools

In this subsection, we collect some preliminary results, which will be used in our proof. Recall (4.1) and (4.7). Denote their resolvents as

$$G_k = (\mathcal{Q}_k - z)^{-1},\ \widetilde{G}_k = (\widetilde{\mathcal{Q}}_k - z)^{-1},\ k = 1, 2. \tag{C.1}$$

We will use the following linearization. For simplicity, let $Y = \Sigma_0^{1/2} X$. Denote the $(N + M) \times (N + M)$ linearized matrix H by

$$H = H(z, X) := \sqrt{z} \begin{pmatrix} 0 & Y \\ Y^* & 0 \end{pmatrix}. \tag{C.2}$$

Similarly, we can define $\widetilde{H}$ by replacing Σ_0 with Σ. By Schur's complement, we have that

$$G(z) = G(z, X) := (H - z)^{-1} = \begin{pmatrix} G_1(z) & \frac{1}{\sqrt{z}} G_1(z) Y \\ \frac{1}{\sqrt{z}} Y^* G_1(z) & G_2(z) \end{pmatrix}. \tag{C.3}$$

The resolvents and related quantities are very convenient for us to analyze the VESD and ESD. Recall the notation in Section 4.3 and the ESD of $\mathcal{Q}_2$ is

$$\zeta = \zeta_N := \frac{1}{M} \sum_{i=1}^{M} \delta_{\lambda_i(\mathcal{Q}_2)}.$$

Denote m_N and $m_{N,\mathbf{b}}$ as the Stieltjes transforms of ζ and ν, respectively. Then we have that

$$m_N = \frac{1}{M} \operatorname{Tr} G_2(z),\ m_{N,\mathbf{b}} = \mathbf{b}^* G_1(z) \mathbf{b}. \tag{C.4}$$

First, we state the anisotropic laws in the following lemma. Recall (4.13). Define the deterministic matrix

$$\Pi(z) := \begin{pmatrix} \Pi_1(z) & 0 \\ 0 & \Pi_2(z) \end{pmatrix} := \begin{pmatrix} -\frac{1}{z}(1 + m(z)\Sigma_0)^{-1} & 0 \\ 0 & m(z) \end{pmatrix}. \tag{C.5}$$

Fix some small constant $\tau > 0$, denote the set of spectral parameters as

$$\mathcal{D} = \mathcal{D}(\tau) = \left\{ z = E + \mathrm{i}\eta : |z| \geq \tau,\ M^{-1+\tau} \leq \eta \leq \tau^{-1} \right\}. \tag{C.6}$$

Moreover, we denote a subset of $\mathcal{D}$ as

$$\mathcal{D}_o = \mathcal{D}_o(\tau) = \mathcal{D} \cap \left\{ \operatorname{dist}(E, \operatorname{supp}(\varrho)) \geq M^{-2/3+\tau} \right\}, \tag{C.7}$$

and the control parameter as

$$\Psi(z) := \sqrt{\frac{\operatorname{Im} m(z)}{M\eta}} + \mathbf{1}(z \in \mathcal{D} \backslash \mathcal{D}_o) \frac{1}{M\eta}.$$

Lemma C.1 (Anisotropic local law). For any deterministic unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{M+N}$, we have that for all $z \in \mathcal{D}(\tau)$

$$\left|\mathbf{u}^* G(z)\mathbf{v} - \mathbf{u}^* \Pi(z)\mathbf{v}\right| \prec \Psi(z).$$

Moreover, we have for all $z \in \mathcal{D}(\tau)$

$$|m_N(z) - m(z)| \prec \frac{1}{N\eta}.$$

Furthermore, when $z \in \mathcal{D}_0(z, \tau)$, we have that

$$|m_N(z) - m(z)| \prec \frac{1}{N(\kappa + \eta)}.$$

Proof. See [51]. ■

We point out that $\operatorname{Im} m(z)$ can be controlled in the following way. Recall ϱ is the measure associated with $m(z)$. We have that

$$\operatorname{Im} m(z) \asymp \begin{cases} \sqrt{\kappa + \eta}, & \text{if } E \in \operatorname{supp} \varrho \\ \frac{\eta}{\sqrt{\kappa+\eta}}, & \text{Otherwise} \end{cases},$$

where $\kappa := \operatorname{dist}(E, \operatorname{supp} \varrho)$. Moreover, according to [76, (4.15) and (4.16)], we have that for $z \in \mathcal{D}$

$$|m(z)| = O(1), \; |m'(z)| = O\left(\frac{1}{\sqrt{\kappa + \eta}}\right). \tag{C.8}$$

Throughout this section, for simplicity of notation, we define the index sets $\mathcal{I}_1 := \{1, 2, \cdots, N\}$, $\mathcal{I}_2 := \{N+1, \cdots, N+M\}$, $\mathcal{I} := \mathcal{I}_1 \cup \mathcal{I}_2$. We shall consistently use the Latin letters $i, j \in \mathcal{I}_1$, Greek letters $\mu, \nu \in \mathcal{I}_2$, and $\mathsf{a}, \mathsf{b} \in \mathcal{I}$. Then we can label the indices of X as $X = (X_{i\mu} : i \in \mathcal{I}_1, \mu \in \mathcal{I}_2)$. For simplicity, given a vector $\mathbf{v} \in \mathcal{C}^{\mathcal{I}_{1,2}}$, we always identity it with its natural embedding in $\mathbb{C}^{\mathcal{I}}$. For example, we shall identify $\mathbf{x} \in \mathbb{C}^{\mathcal{I}_1}$ with $\begin{pmatrix} \mathbf{x} \\ \mathbf{0} \end{pmatrix}$, and $\mathbf{y} \in \mathbb{C}^{\mathcal{I}_2}$ with $\begin{pmatrix} \mathbf{0} \\ \mathbf{y} \end{pmatrix}$. We will also consistently use the notation $G_{\mathbf{xy}}(z) = \mathbf{x}^* G(z)\mathbf{y}$. Second, we will frequently use the following identities.

Lemma C.2 (Ward's identity). Let $\{\mathbf{u}_i\}_{i\in\mathcal{I}_1}$ and $\{\mathbf{v}_\mu\}_{\mu\in\mathcal{I}_2}$ be orthonormal basis vectors in $\mathbb{R}^{\mathcal{I}_1}$ and $\mathbb{R}^{\mathcal{I}_2}$, respectively. For $\mathbf{x}\in\mathbb{C}^{\mathcal{I}_1}$ and $\mathbf{y}\in\mathbb{C}^{\mathcal{I}_2}$, we have

$$\sum_{i\in\mathcal{I}_1}|G_{\mathbf{x}\mathbf{u}_i}|^2=\sum_{i\in\mathcal{I}_1}|G_{\mathbf{u}_i\mathbf{x}}|^2=\frac{|z|^2}{\eta}\mathrm{Im}\left(\frac{G_{\mathbf{xx}}}{z}\right),$$

$$\sum_{\mu\in\mathcal{I}_2}|G_{\mathbf{y}\mathbf{v}_\mu}|^2=\sum_{\mu\in\mathcal{I}_2}|G_{\mathbf{v}_\mu\mathbf{y}}|^2=\frac{\mathrm{Im}\,G_{\mathbf{yy}}(z)}{\eta},$$

$$\sum_{i\in\mathcal{I}_1}|G_{\mathbf{y}\mathbf{u}_i}|^2=\sum_{i\in\mathcal{I}_1}|G_{\mathbf{u}_i\mathbf{y}}|^2=G_{\mathbf{yy}}+\frac{\bar{z}}{\eta}\mathrm{Im}\,G_{\mathbf{yy}},$$

$$\sum_{\mu\in\mathcal{I}_2}|G_{\mathbf{x}\mathbf{v}_\mu}|^2=\sum_{\mu\in\mathcal{I}_2}|G_{\mathbf{v}_\mu\mathbf{x}}|^2=\frac{G_{\mathbf{xx}}}{z}+\frac{\bar{z}}{\eta}\mathrm{Im}\left(\frac{G_{\mathbf{xx}}}{z}\right).$$

Proof. The proofs follow from the spectral decomposition of G as in (C.3) and the orthonormality of the basis. See Lemma 4.1 of [76] for details. ■

Third, we will also need the following estimate.

Lemma C.3. For any two vectors $\mathbf{b}_1,\mathbf{b}_2\in\mathbb{R}^{\mathcal{I}}$, we have that

$$\sum_{\mu\in\mathcal{I}_2}G_{\mathbf{b}_1\mu}(z_1)G_{\mathbf{b}_2\mu}(z_2)=\frac{\Pi_{\mathbf{b}_1\mathbf{b}_2}(z_1)-\Pi_{\mathbf{b}_1\mathbf{b}_2}(z_2)}{z_1-z_2}+\mathrm{O}_\prec(\eta^{-1}(N\eta)^{-1/2}),$$

where we used the convention that

$$\lim_{z_2\to z_1}\frac{\Pi_{\mathbf{b}_1\mathbf{b}_2}(z_1)-\Pi_{\mathbf{b}_1\mathbf{b}_2}(z_2)}{z_1-z_2}=\Pi'_{\mathbf{b}_1\mathbf{b}_2}(z_1).$$

Proof. Note that by spectral decomposition, we have that

$$\sum_{\mu\in\mathcal{I}_2}G_{\mathbf{b}_1\mu}(z_1)G_{\mathbf{b}_2\mu}(z_2)=\frac{G_{\mathbf{b}_1\mathbf{b}_2}(z_1)-G_{\mathbf{b}_1\mathbf{b}_2}(z_2)}{z_1-z_2}. \tag{C.9}$$

The proof follows from local law and Cauchy's integral formula. See equation (5.21) of [76] for more details. ■

Finally, we introduce the device of cumulant expansion. Recall that for any random variable h its kth cumulant is defined as

$$\kappa_k(h)=\left(\partial_t^k\log\mathbb{E}e^{th}\right)|_{t=0}. \tag{C.10}$$

Lemma C.4 (Cumulant expansion). Fix any $\ell \in \mathbb{N}$ and let $f \in \mathrm{C}^{\ell+1}(\mathbb{R})$ be a complex-valued function. Suppose h is a real valued random variable with finite moments up to order $\ell+2$. Then we have that

$$\mathbb{E}(f(h)h) = \sum_{k=0}^{\ell} \frac{1}{k!} \kappa_{k+1}(h) \mathbb{E} f^{(k)}(h) + R_{\ell+1},$$

where $\kappa_k(h)$ is the kth cumulant of h and $\mathbb{R}_{\ell+1}$ satisfies

$$R_{\ell+1} \lesssim \mathbb{E}\left| h^{\ell+2} \mathbf{1}_{|h|>N^{-1/2+\epsilon}} \right| \cdot \|f^{(\ell+1)}\|_{\infty} + \mathbb{E}|h|^{\ell+2} \cdot \sup_{|x| \leq N^{-1/2+\epsilon}} |f^{(\ell+1)}(x)|,$$

for any constant $\epsilon > 0$.

Proof. See [60, Proposition 3.1] or [50, Section II]. ■

C.2 The non-spiked case: CLT for $\mathcal{Y}$

In this subsection, we prove the CLT for $\mathcal{Y}$ defined in (4.12). Note that we can write the integrand into a trace form. As before, we set the natural embedding of $\mathbf{b} \in \mathbb{R}^N$ as $\mathbf{b} \in \mathbb{R}^{N+M}$ such that

$$\mathbf{b} = \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix}. \tag{C.11}$$

Additionally, we denote $B = \mathbf{b}\mathbf{b}^*$. According to (C.4) and (4.9), we can write

$$\mathcal{Y} = \sqrt{M\eta} \oint_{\Gamma} \mathsf{g}(z) \operatorname{Tr}([G(z) - \Pi(z)]B) \mathrm{d}z, \ \mathsf{g}(z) = \frac{g(z)}{2\pi \mathrm{i}}. \tag{C.12}$$

Recall that if x is a real Gaussian random variable, that is, $x \sim \mathcal{N}(0, \sigma^2)$, denote $\mathbbm{m}_n = \mathbb{E}x^n$, we have that

$$\mathbbm{m}_{n+2} = (n+1)\sigma^2 \mathbbm{m}_n. \tag{C.13}$$

Our goal is to prove an asymptotic version of (C.13) for $\mathcal{Y}$.

For $\Pi(z)$ defined in (C.5), we introduce the following auxiliary quantities for the ease of statements

$$A_1 = -z\Pi(z), \ A_2 = I - A_1. \tag{C.14}$$

In view of (C.3), we will frequently use the following identity:

$$G = \frac{1}{z}(HG - I). \tag{C.15}$$

The starting point is to decompose the following quantity $\mathcal{Z} := \sqrt{M\eta}\,\mathrm{Tr}([G(z)-\Pi(z)]B)$, so that

$$\begin{aligned}\mathcal{Z} &= \sqrt{M\eta}\left(\mathrm{Tr}(GBA_1) - \mathrm{Tr}(\Pi B) + \mathrm{Tr}(GBA_2)\right)\\ &= \sqrt{M\eta}\left(\frac{1}{z}\,\mathrm{Tr}(HGBA_1) - \frac{1}{z}\,\mathrm{Tr}\,BA_1 - \mathrm{Tr}(\Pi B) + \mathrm{Tr}(GBA_2)\right)\\ &= \sqrt{M\eta}\left(\frac{1}{z}\,\mathrm{Tr}(HGBA_1) + \mathrm{Tr}(GBA_2)\right), \end{aligned} \tag{C.16}$$

where in the second step we used (C.15) and in the third step we used the definition of A_1 as in (C.14). Together with (C.12), for any integer k, we have that

$$\begin{aligned}\mathbb{E}\mathcal{Y}^k = \sqrt{M\eta}&\left[\mathbb{E}\oint_\Gamma \frac{\mathsf{g}(z)}{z}\,\mathrm{Tr}(HGBA_1)\mathrm{d}z\mathcal{Y}^{k-1}\right]\\ &+\sqrt{M\eta}\left[\mathbb{E}\oint_\Gamma \mathsf{g}(z)\,\mathrm{Tr}(GBA_2)\mathrm{d}z\mathcal{Y}^{k-1}\right].\end{aligned} \tag{C.17}$$

Denote $j' = j+N$ and $\Lambda \in \mathbb{R}^{(M+N)\times(M+N)}$ as

$$\Lambda := \begin{pmatrix}\Sigma_0^{1/2} & 0\\ 0 & I\end{pmatrix}. \tag{C.18}$$

We have that

$$\mathrm{Tr}\,HGBA_1 = \sqrt{z}\sum_{i,j} X_{ij}(GBA_1\Lambda)_{ji'}. \tag{C.19}$$

Let $E_{ij'}$ be an $(M+N)\times(M+N)$ matrix whose only nonzero entry is the (i,j')th entry and equals to one. We next prepare some expressions for derivatives that follow from elementary calculations. Note that

$$\frac{\partial G}{\partial X_{ij}} = -G\frac{\partial H}{\partial X_{ij}}G = -\sqrt{z}G(\Lambda E_{ij'} + E_{j'i}\Lambda)G. \tag{C.20}$$

Consequently, for any block diagonal matrix D, we have that

$$\left(\frac{\partial G}{\partial X_{ij}}BD\right)_{j'i} = -\sqrt{z}\left[(G\Lambda)_{j'i}(GBD)_{j'i} + G_{j'j'}(\Lambda GBD)_{ii}\right]. \tag{C.21}$$

Additionally, we have that

$$\begin{aligned}\frac{\partial \mathcal{Z}}{\partial X_{ij}} &= -\sqrt{zM\eta}\,\mathrm{Tr}\left(G(\Lambda E_{ij'} + E_{j'i}\Lambda)GB\right)\\ &= -\sqrt{zM\eta}\left[(GBG\Lambda)_{ij'} + (\Lambda GBG)_{j'i}\right].\end{aligned} \tag{C.22}$$

From now on, we will conduct calculations on (C.17). Our strategy is to focus on the first term of the right-hand side of (C.17) as we will see later that that second term will be canceled algebraically. Denote

$$h_1 = h_1(i,j) := (GBA_1\Lambda)_{j'i},\ h_2 = \mathcal{Y}^{k-1}. \tag{C.23}$$

Since $\mathsf{g}(z)$ is purely deterministic, using Lemma C.4, (C.19), and (C.21), we readily obtain

$$\sqrt{M\eta}\mathbb{E}\oint_\Gamma \frac{\mathsf{g}(z)}{z}\operatorname{Tr}(HGBA_1)\mathrm{d}z\mathcal{Y}^{k-1} = \mathbb{E}(P_1+P_2+P_3).$$

Here P_1 is defined as

$$P_1 := \oint_\Gamma \mathsf{g}(z)\left(-\frac{\sqrt{\eta}}{\sqrt{M}}\sum_{i,j}(G\Lambda)_{j'i}(GBA_1\Lambda)_{j'i} - \frac{\sqrt{\eta}}{\sqrt{M}}\sum_{i,j}G_{j'j'}(\Lambda GBA_1\Lambda)_{ii}\right)\mathrm{d}zh_2,$$

and P_2 is defined as

$$P_2 := \frac{\sqrt{\eta}}{\sqrt{M}}\oint_\Gamma \sqrt{z}\mathsf{g}(z)\sum_{i,j}\left(GBA_1\Lambda\right)_{j'i}\mathrm{d}z\frac{\partial h_2}{\partial X_{ij}}, \tag{C.24}$$

and P_3 is defined as

$$\begin{aligned} P_3 :&= \sqrt{\eta}\sum_{l=2}^{3}\frac{\kappa_{l+1}}{l!\,M^{l/2}}\sum_{i,j}\frac{\partial^l}{\partial X_{ij}^l}(\oint_\Gamma \mathsf{g}(z)h_1\mathrm{d}zh_2) + R_1 \\ &:= P_{31}+P_{32}+R_1. \end{aligned} \tag{C.25}$$

In the last equation, P_{31} collects the summation for $l=2$, P_{32} collects that of the summation for $l=3$, and $R_1 := P_3 - P_{31} - P_{32}$ is the residual. Here we used the notation that

$$\frac{\partial^l}{\partial X_{ij}^l}(h_1h_2) = \sum_{l_1+l_2=l}\binom{l}{l_1,l_2}\frac{\partial^{l_1}h_1}{\partial X_{ij}^{l_1}}\frac{\partial^{l_2}h_2}{\partial X_{ij}^{l_2}},\quad \binom{l}{l_1,l_2} = \frac{l!}{l_1!\,l_2!}.$$

We will see later that $l=2$ will contribute nothing, $l=3$ will give some extra terms that explains the fourth moment contributes, and $l=4$ is needed to show R_1 is small.

For P_1, on one hand, using (C.3), by definitions of A_1 and Λ, we have that

$$\begin{aligned} \frac{\sqrt{\eta}}{\sqrt{M}}\sum_{i,j}(G\Lambda)_{j'i}(GBA_1\Lambda)_{j'i} &\asymp \frac{\sqrt{\eta}}{\sqrt{M}}\sum_{i,j}(Y^*G_1\Sigma_0^{1/2})_{ji}(Y^*G_1\mathbf{b}\mathbf{b}^*\Pi\Sigma_0^{1/2})_{ji} \\ &= \frac{\sqrt{\eta}}{\sqrt{M}}\operatorname{Tr}\Sigma_0^{1/2}G_1YY^*G_1\mathbf{b}\mathbf{b}^*\Pi_1\Sigma_0^{1/2} \\ &= \frac{\sqrt{\eta}}{\sqrt{M}}\mathbf{b}_1^*G_1YY^*G_1\mathbf{b}, \end{aligned}$$

where we denote $\mathbf{b}_1 := \Pi_1\Sigma_0\mathbf{b}$. In view of (C.3), we have that for $z\in\Gamma$

$$\frac{\sqrt{\eta}}{\sqrt{M}}\sum_{i,j}(G\Lambda)_{j'i}(GBA_1\Lambda)_{j'i} \asymp \frac{\sqrt{\eta}}{\sqrt{M}}\sum_{\mu\in\mathcal{I}_2}\mathbf{b}_1^*G\mathbf{e}_\mu\mathbf{b}^*G\mathbf{e}_\mu \prec \frac{1}{\sqrt{M}},$$

where we used Lemma C.1, the definition of Γ, and (C.8) in the last step. On the other hand, using Lemma C.1, we have that

$$-\frac{\sqrt{\eta}}{\sqrt{M}}\sum_{i,j} G_{j'j'}(\Lambda GBA_1\Lambda)_{ii} = -\sqrt{M\eta}\,\mathrm{Tr}\, G_2\,\mathrm{Tr}(GBA_1\Lambda^2)$$

$$= -m(z)\sqrt{M\eta}\,\mathrm{Tr}\, GBA_1\Lambda^2 + \mathrm{O}_{\prec}\left((M\eta)^{-1/2}\right)$$

$$= -\sqrt{M\eta}\,\mathrm{Tr}\, GBA_2 + \mathrm{O}_{\prec}\left((M\eta)^{-1/2}\right),$$

where in the last step we used the fact that $m(z)BA_1\Lambda^2 = BA_2$, which follows directly from (C.14).

Using the above calculations and inserting them back into (C.17), we find that

$$\mathbb{E}\mathcal{Y}^k = \mathbb{E}P_2 + \mathbb{E}P_3 + \mathrm{O}_{\prec}((M\eta)^{-1/2}). \tag{C.26}$$

We summarize the properties of P_2 and P_3 in the following lemma and defer its proof to Sections C.3 and C.4.

Lemma C.5. We have that

$$P_2 = (k-1)\mathsf{V}_1(\mathbf{b},\mathbf{b})\mathcal{Y}^{k-2} + \mathrm{O}_{\prec}((N\eta)^{-1/2}), \tag{C.27}$$

and

$$P_3 = (k-1)\kappa_4\mathsf{V}_2(\mathbf{b},\mathbf{b})\mathcal{Y}^{k-2} + \mathrm{O}_{\prec}((N\eta)^{-1/2}). \tag{C.28}$$

Recall (C.13). It is easy to see that Theorem 4.3 follows from Lemma C.5 and (C.26).

C.3 Proof of Lemma C.5: Verification of (C27)

We first provide some useful results. By Lemma C.1, it is easy to see that

$$\mathcal{Y} = \mathrm{O}_{\prec}(1). \tag{C.29}$$

Moreover, using the definition of $\mathcal{Y}$ and (C.22), we have

$$\frac{\partial h_2}{\partial X_{ij}} = (k-1)\mathcal{Y}^{k-2}\oint_\Gamma \mathsf{g}(z)\frac{\partial \mathcal{Z}}{\partial X_{ij}}\mathrm{d}z$$

$$= (k-1)\mathcal{Y}^{k-2}\oint_\Gamma \mathsf{g}(z)\left(-\sqrt{zM\eta}\left[(GBG\Lambda)_{ij'} + (\Lambda GBG)_{j'i}\right]\right)\mathrm{d}z. \tag{C.30}$$

Consequently, in view of (C.24), we can write

$$P_2 = -(k-1)\eta\mathbb{E}\mathcal{L}\mathcal{Y}^{k-2}, \tag{C.31}$$

where $\mathcal{L}$ is defined as

$$\mathcal{L} := 2\sum_{i,j}\oint_\Gamma\oint_\Gamma \sqrt{z_1z_2}\mathsf{g}(z_1)\mathsf{g}(z_2)(G(z_1)BA_1(z_1)\Lambda)_{j'i}(G(z_2)BG(z_2)\Lambda)_{ij'}\mathrm{d}z_1\mathrm{d}z_2$$
$$= 2\oint_\Gamma\oint_\Gamma \mathsf{g}(z_1)\mathsf{g}(z_2)\,\mathrm{Tr}(\Sigma_0^{1/2}G_1(z_2)\mathbf{b}\mathbf{b}^*G_1(z_2)YY^*G_1(z_1)\mathbf{b}\mathbf{b}^*(1+m(z_1)\Sigma_0)^{-1}\Sigma_0^{1/2})\mathrm{d}z_1\mathrm{d}z_2$$
$$= 2\oint_\Gamma\oint_\Gamma \mathsf{g}(z_1)\mathsf{g}(z_2)\left[\mathbf{b}^*(1+m(z_1)\Sigma_0)^{-1}\Sigma_0G_1(z_2)\mathbf{b}\right]\left[\mathbf{b}^*G_1(z_2)YY^*G_1(z_1)\mathbf{b}\right]\mathrm{d}z_1\mathrm{d}z_2. \tag{C.32}$$

Using the structure of (C.3) and (C.11), we have that

$$\mathbf{b}^*G_1(z_2)YY^*G_1(z_1)\mathbf{b} = \sqrt{z_1z_2}\sum_{\mu\in\mathcal{I}_2}\mathbf{b}^*G\mathbf{e}_\mu\mathbf{b}^*G\mathbf{e}_\mu$$
$$= \sqrt{z_1z_2}\frac{\mathbf{b}^*(\Pi_1(z_1)-\Pi_1(z_2))\mathbf{b}}{z_1-z_2} + \mathrm{O}_\prec(\eta^{-1}(N\eta)^{-1/2}), \tag{C.33}$$

where in the last step we used Lemma C.3. The rest of the proof follows from Lemma C.1 and (C.29).

C.4 Proof of Lemma C.5: verification of (C28)

To control P_3, we separate our discussion in the following three subsections according to the order of the expansion as in (C.25).

C.4.1 $l = 2$

This corresponds to the term P_{31} in (C.25). Formally, we can write

$$P_{31} = \frac{\kappa_3}{2}\frac{\sqrt{\eta}}{M}\mathbb{E}\sum_{i,j}\left(\mathsf{P}_{31}(2,0)+\mathsf{P}_{31}(1,1)+\mathsf{P}_{31}(0,2)\right),$$

where we denote

$$\mathsf{P}_{31}(2,0) = \mathsf{P}_{31}(2,0;i,j) =: \oint_\Gamma \mathsf{g}(z)\left(\frac{\partial^2 G}{\partial X_{ij}^2}BA_1\Lambda\right)_{j'i}\mathrm{d}zh_2, \tag{C.34}$$

$$\mathsf{P}_{31}(1,1) = \mathsf{P}_{31}(1,1;i,j) = 2\oint_\Gamma \mathsf{g}(z)\left(\frac{\partial G}{\partial X_{ij}}BA_1\Lambda\right)_{j'i}\mathrm{d}z\frac{h_2}{\partial X_{ij}}, \tag{C.35}$$

$$\mathsf{P}_{31}(0,2) = \mathsf{P}_{31}(0,2;i,j) = \oint_\Gamma \mathsf{g}(z)\left(GBA_1\Lambda\right)_{j'i}\mathrm{d}z\frac{\partial^2 h_2}{\partial X_{ij}^2}. \tag{C.36}$$

We first prepare some useful identities, which can be obtained using some elementary calculation. Using (C.22) and (C.20), we have that

$$\frac{\partial^2 \mathcal{Z}}{\partial X_{ij}^2} = z\sqrt{M}\eta\Big((G\Lambda)_{ii}(GBG\Lambda)_{j'j'} + G_{ij'}(\Lambda GBG\Lambda)_{ij'} + (GBG\Lambda)_{ii}(G\Lambda)_{j'j'} + (GBG)_{ij'}(\Lambda G\Lambda)_{ij'} + (\Lambda G\Lambda)_{j'i}(GBG)_{j'i} + (\Lambda G)_{j'j'}(\Lambda GBG)_{ii} + (\Lambda GBG\Lambda)_{j'i}G_{j'i} + (\Lambda GBG)_{j'j'}(\Lambda G)_{ii}\Big). \tag{C.37}$$

Moreover, we have that

$$\frac{\partial^2 h_2}{\partial X_{ij}^2} = (k-1)(k-2)\mathcal{Y}^{k-3}\left(\oint_\Gamma \mathsf{g}(z)\frac{\partial \mathcal{Z}}{\partial X_{ij}}\mathrm{d}z\right)^2 + (k-1)\mathcal{Y}^{k-2}\oint_\Gamma \mathsf{g}(z)\frac{\partial^2 \mathcal{Z}}{\partial X_{ij}^2}\mathrm{d}z. \tag{C.38}$$

Additionally, we have

$$\frac{\partial^2 G}{\partial X_{ij}^2} = 2z\left[G(\Lambda E_{ij'} + E_{j'i}\Lambda)\right]^2 G.$$

For the ease of discussion, in what follows, we use the following shorthand notation

$$\mathcal{L}_{ij} := \Lambda E_{ij'} + E_{j'i}\Lambda. \tag{C.39}$$

For any block-diagonal matrix $D = D_1 \oplus D_2$, we have that

$$\begin{aligned}\left(\frac{\partial^2 G}{\partial X_{ij}^2}BD\right)_{j'i} &= 2z\left(G\mathcal{L}_{ij}G\mathcal{L}_{ij}GBD\right)_{j'i} \\ &= 2z\left[(G\mathcal{L}_{ij}G)_{j'j'}(GBD)_{ii} + (G\mathcal{L}_{ij}G)_{j'i}(GBD)_{j'i}\right].\end{aligned} \tag{C.40}$$

Note that

$$(G\mathcal{L}_{ij}G)_{j'j'} = 2(G\Lambda)_{j'i}G_{j'j'},\ \ (GBD)_{ii} = (G_1\mathbf{bb}^*D_1)_{ii},\ (G\mathcal{L}_{ij}G)_{ii} = (G\Lambda)_{ii}G_{j'i} + G_{ij'}(\Lambda G)_{ii}. \tag{C.41}$$

and

$$(G\mathcal{L}_{ij}G)_{j'i} = (G\Lambda)_{j'i}G_{ii} + G_{j'i}(\Lambda G)_{ii}, \tag{C.42}$$

$$(GBD)_{j'i} = (G_{21}\mathbf{bb}^*D_1)_{ji}. \tag{C.43}$$

Moreover, we have that

$$(\Lambda GBG)_{ii} = (\Sigma^{1/2}G_1\mathbf{bb}^*G_1)_{ii},\ (\Lambda GBG)_{ij'} = (\Sigma_0^{1/2}G_1\mathbf{bb}^*G_{12})_{ij}, \tag{C.44}$$

and

$$(\Lambda GBG)_{j'j'} = (G_{21}\mathbf{b}\mathbf{b}^* G_{12})_{j'j'}, \tag{C.45}$$

where we used the conventions that $G_{12} = z^{-1/2}G_1 Y$ and $G_{21} = G_{12}^*$.

We give a more explicitly form of D_1. As in (C.19), $D = A_1\Lambda$. Consequently, we shall have that

$$D_1 = -z\Pi_1(z)\Sigma_0^{1/2}. \tag{C.46}$$

This leads to that

$$\mathbf{b}^* D_1 \mathbf{f}_i = -z\mathbf{b}^*\Pi_1(z)\Sigma_0^{1/2}\mathbf{f}_i = \mathbf{f}_i^*\Sigma^{1/2}\Pi_1(z)\mathbf{b}. \tag{C.47}$$

Note that D_1 is symmetric since Π_1 and Σ_0 share the same eigenvectors.

We summarize the main estimates in the following lemma.

Lemma C.6. We have the following estimates:

$$\frac{\kappa_3}{2}\frac{\sqrt{\eta}}{M}\sum_{i,j}\mathsf{P}_{31}(2,0) = \mathrm{O}_{\prec}\left(\frac{1}{\sqrt{M\eta}}\right), \tag{C.48}$$

$$\frac{\kappa_3}{2}\frac{\sqrt{\eta}}{M}\sum_{i,j}\mathsf{P}_{31}(1,1) = \mathrm{O}_{\prec}\left(\frac{1}{\sqrt{M\eta}}\right), \tag{C.49}$$

$$\frac{\kappa_3}{2}\frac{\sqrt{\eta}}{M}\sum_{i,j}\mathsf{P}_{31}(0,2) = \mathrm{O}_{\prec}\left(\frac{1}{\sqrt{M\eta}}\right). \tag{C.50}$$

Proof. (1). Justification of (C.48). In view of (C.40), (C.41), and the definition of P_{31}, we focus our discussion on some typical terms. By Lemma C.1, we see that

$$\begin{aligned}\left|\frac{\sqrt{\eta}}{M}\sum_{i,j} G_{j'j'}(G\Lambda)_{j'i}(G_1\mathbf{b}\mathbf{b}^* D_1)_{ii}\right| &\prec \frac{1}{\sqrt{\eta}}\frac{\sqrt{\eta}}{M^{3/2}}\sum_{i,j}|\mathbf{f}_i^* G_1\mathbf{b}||\mathbf{f}_i^* D_1\mathbf{b}| \\ &\prec \frac{1}{\sqrt{M}}\sum_i |\mathbf{f}_i^* G_1\mathbf{b}||\mathbf{f}_i^* D_1\mathbf{b}|,\end{aligned}$$

where in the first step we used $(\Pi\Lambda)_{j'i} = 0$ and the symmetry of D_1. Applying the Cauchy–Schwarz inequality, we have that

$$\begin{aligned}\sum_i |\mathbf{f}_i^* G_1\mathbf{b}||\mathbf{f}_i^* D_1\mathbf{b}| &\le \left(\sum_i |\mathbf{f}_i^* D_1\mathbf{b}|^2\right)^{1/2}\left(\sum_i |\mathbf{f}_i^* G_1\mathbf{b}|^2\right)^{1/2} \\ &= \left(\sum_i |\mathbf{f}_i^* D_1\mathbf{b}|^2\right)^{1/2}\left(\sum_i |G_{\mathbf{e}_i\mathbf{b}}|^2\right)^{1/2}.\end{aligned} \tag{C.51}$$

Using (C.47), it is easy to see that

$$\sum_i |\mathbf{f}_i^* D_1 \mathbf{b}|^2 \asymp \sum_i \mathbf{f}_i^* \Sigma_0^{1/2} \Pi_1(z) \mathbf{b}\mathbf{b}^* \Pi_1 \Sigma_0^{1/2} \mathbf{f}_i = \mathbf{b}^* \Pi_1 \Sigma_0 \Pi_1 \mathbf{b} \asymp 1. \tag{C.52}$$

Inserting the above estimate back into (C.51), together with Ward's identities in Lemma C.2, we find that

$$\sum_i |\mathbf{f}_i^* G \mathbf{b}||\mathbf{f}_i^* D \mathbf{b}| \prec \frac{1}{\sqrt{\eta}}, \tag{C.53}$$

where we used Lemma C.1 and (C.8) to obtain that $\operatorname{Im}(G_{\mathbf{b}\mathbf{b}}/z) \asymp 1$. This yields that

$$\left| \frac{\sqrt{\eta}}{M} \sum_{i,j} G_{j'j'} (G\Lambda)_{j'i} (G_1 \mathbf{b}\mathbf{b}^* D_1)_{ii} \right| = \mathrm{O}_{\prec}\left(\frac{1}{\sqrt{M\eta}} \right).$$

Similarly, we have that

$$\begin{aligned} \left| \frac{\sqrt{\eta}}{M} \sum_{i,j} (G\Lambda)_{j'i} G_{ii} (Y^* G_1 \mathbf{b}\mathbf{b}^* D_1)_{ji} \right| &\prec \frac{1}{M^{3/2}} \sum_{i,j} |\mathbf{f}_j^* Y^* G_1 \mathbf{b}^*||\mathbf{b}^* D_1 \mathbf{f}_i| \\ &\prec \frac{1}{M\sqrt{\eta}} \sum_i |\mathbf{b}^* D \mathbf{f}_i| \prec \frac{1}{\sqrt{M\eta}}, \end{aligned}$$

where in the second step we used Lemma C.1 and in the last step we used the Cauchy–Schwarz inequality and (C.52) to obtain that for some constant $C > 0$

$$\sum_i |\mathbf{b}^* D_1 \mathbf{f}_i| \leq \sqrt{N} \left(\sum_i |\mathbf{b}^* D_1 \mathbf{f}_i|^2 \right) \leq C\sqrt{N}. \tag{C.54}$$

Analogously, we can show that

$$\left| \frac{\sqrt{\eta}}{M} \sum_{i,j} G_{j'i} (\Lambda G)_{ii} (Y^* G_1 \mathbf{b}\mathbf{b}^* D_1)_{ji} \right| \prec \frac{1}{\sqrt{M\eta}}.$$

Using (C.40) and the formulas below, in view of the definition (C.34), combing the above estimates and (C.29), we have concluded our proof.

(2). Justification of (C.49). We again work with some typical terms. Set

$$\mathbf{v}_i = \Sigma_0^{1/2} \mathbf{f}_i. \tag{C.55}$$

By Lemma C.1, we have that

$$\left|\sqrt{M\eta}\frac{\sqrt{\eta}}{M}\sum_{i,j} G_{j'j'}(z_1)(\Lambda G(z_1)BD(z_1))_{ii}(G(z_2)BG(z_2)\Lambda)_{ij'}\right|$$

$$\prec \frac{\eta}{\sqrt{M}}\sum_{i,j}(\Sigma_0^{1/2}G_1(z_1)\mathbf{b}\mathbf{b}^*D_1(z_1))_{ii}(G_1(z_2)\mathbf{b}\mathbf{b}^*G_1(z_2)Y)_{ij}+\frac{1}{M^{3/2}}\sum_{i,j}|\mathbf{b}^*G_1(z_1)\mathbf{v}_i\mathbf{b}^*D_1(z_1)\mathbf{f}_i|$$

$$\prec \frac{\eta}{\sqrt{M}}\sum_{i,j}\mathbf{b}^*G_1(z_1)\mathbf{v}_i\mathbf{b}^*D_1(z_1)\mathbf{f}_i\mathbf{b}^*G_1(z_2)Y\mathbf{f}_j+\frac{1}{\sqrt{M}}\sum_{i}|\mathbf{b}^*G_1(z_1)\mathbf{v}_i\mathbf{b}^*D_1(z_1)\mathbf{f}_i|.$$

We first consider the second term of the right-hand side of the above equation. The discussion is similar to (C.51) and (C.53) except that $\{\mathbf{v}_i\}$ may not be an orthonormal basis so that Lemma C.2 cannot be applied directly. Note that since $\|\Sigma_0\|$ is bounded, by the Cauchy-Schwarz inequality, we have that for some constant $C>0$

$$\sum_i|\mathbf{b}^*G_1(z_1)\mathbf{v}_i|^2=\mathbf{b}^*G_1(z_1)\Sigma_0\overline{G}_1(z_1)\mathbf{b}\leq C\sum_i|\mathbf{b}^*G_1(z_1)\mathbf{f}_i|^2. \tag{C.56}$$

As a result, together with (C.53), we readily obtain that

$$\frac{1}{\sqrt{M}}\sum_i|\mathbf{b}^*G_1(z_1)\mathbf{v}_i\mathbf{b}^*D_1(z_1)\mathbf{f}_i|\prec\frac{1}{\sqrt{M\eta}}. \tag{C.57}$$

The first term can be controlled similarly using that

$$\sum_j \mathbf{b}^*G_1(z_2)Y\mathbf{f}_j=\mathbf{b}^*G_1(z)Y\mathbf{1}\prec\frac{\sqrt{N}}{\sqrt{N\eta}}=\frac{1}{\sqrt{\eta}},$$

where $\mathbf{1}$ is a vector with all unity and in the last step we used Lemma C.1. This yields that

$$\left|\sqrt{M\eta}\frac{\sqrt{\eta}}{M}\sum_{i,j} G_{j'j'}(z_1)(\Lambda G(z_1)BD(z_1))_{ii}(G(z_2)BG(z_2)\Lambda)_{ij'}\right|\prec\frac{1}{\sqrt{M\eta}}.$$

Similarly, we can show that

$$\left|\sqrt{M\eta}\frac{\sqrt{\eta}}{M}\sum_{i,j}(G(z_1)\Lambda)_{j'i}(G(z_1)BD(z_2))_{j'i}(G(z_2)BG(z_2)\Lambda)_{ij'}\right|\prec\frac{1}{\sqrt{M\eta}}.$$

Using(C.22) and (C.21), in view of the definition (C.35), by (C.29), we have completed the proof.

(3). Justification of (C.50). We work on some typical terms according to (C.38). By Lemma C.1, we have that

$$\left|\frac{\sqrt{\eta}}{M}M\eta\sum_{i,j}(G(z_0)BA_1(z_0)\Lambda)_{j'i}(G(z_1)BG(z_1)\Lambda)_{ij'}(G(z_2)BG(z_2)\Lambda)_{ij'}\right|$$

$$\asymp \eta^{3/2}\sum_{i,j}|\mathbf{b}^*G_1(z_1)\mathbf{f}_i||\mathbf{b}^*G_1(z_2)\mathbf{f}_i||\mathbf{f}_j^*Y^*G_1(z_1)\mathbf{b}||\mathbf{f}_j^*Y^*G_1(z_2)\mathbf{b}||\mathbf{f}_j^*Y^*G_1(z_0)\mathbf{b}||\mathbf{b}^*\Pi_1(z_0)\mathbf{v}_i\mathbf{b}|$$

$$\prec \eta^{3/2}\frac{M}{(M\eta)^{3/2}}\sum_{i,j}|\mathbf{b}^*G_1(z_1)\mathbf{f}_i||\mathbf{b}^*G_1(z_2)\mathbf{f}_i||\mathbf{b}^*\Pi_1(z_0)\mathbf{v}_i|$$

$$\prec \frac{1}{\sqrt{M}}\sum_{i}|\mathbf{b}^*G_1(z_2)\mathbf{f}_i||\mathbf{b}^*\Pi_1(z_0)\mathbf{v}_i| \prec \frac{1}{\sqrt{M\eta}},$$

where in the last step we used (C.57). Similarly, we have that

$$\left|\frac{\sqrt{\eta}}{M}\sqrt{M\eta}\sum_{i,j}\left(G(z_1)BA_1(z_1)\Lambda\right)_{j'i}(G(z_2)\Lambda)_{ii}(G(z_2)BG(z_2)\Lambda)_{j'j'}\right|$$

$$\prec \frac{\eta}{\sqrt{M}}\frac{M}{(M\eta)^{3/2}}\sum_{i}|\mathbf{b}^*\Pi_1(z_1)\mathbf{v}_i| \asymp \frac{1}{\sqrt{M\eta}},$$

and

$$\left|\frac{\sqrt{\eta}}{M}\sqrt{M\eta}\sum_{i,j}(G(z_1)BA_1(z_1)\Lambda)_{j'i}(\Lambda G(z_2)\Lambda)_{j'i}(G(z_2)BG(z_2))_{j'i}\right|$$

$$\prec \frac{\eta}{\sqrt{M}}\frac{M}{(M\eta)^{3/2}}\sum_{i}|\mathbf{b}^*\Pi_1(z_1)\mathbf{v}_i| \asymp \frac{1}{\sqrt{M\eta}}.$$

The other terms can be analyzed in the same way. Using (C.38) and (C.37), in view of the definition (C.36), combing the above estimates and (C.29), we have concluded our proof. ■

C.4.2 $l=3$

This corresponds to the term P_{32} in (C.25). We decompose P_{32} as follows:

$$P_{31} = \frac{\kappa_4}{6}\frac{\sqrt{\eta}}{M^{3/2}}\mathbb{E}\sum_{i,j}\left(\mathsf{P}_{32}(1,2)+\mathsf{P}_{32}(2,1)+\mathsf{P}_{32}(0,3)+\mathsf{P}(3,0)\right),$$

where we denote

$$\mathsf{P}_{32}(1,2)=\mathsf{P}_{32}(1,2,i,j):=3\oint_\Gamma \mathsf{g}(z)\left(\frac{\partial G}{\partial X_{ij}}BA_1\Lambda\right)_{j'i}\mathrm{d}z\frac{\partial^2 h_2}{\partial X_{ij}^2},$$
$$\mathsf{P}_{32}(2,1):=3\oint_\Gamma \mathsf{g}(z)\left(\frac{\partial^2 G}{\partial X_{ij}^2}BA_1\Lambda\right)_{j'i}\mathrm{d}z\frac{\partial h_2}{\partial X_{ij}},$$
$$\mathsf{P}_{32}(3,0):=\oint_\Gamma \mathsf{g}(z)\left(\frac{\partial^3 G}{\partial X_{ij}^3}BA_1\Lambda\right)_{j'i}\mathrm{d}z h_2,$$
$$\mathsf{P}_{32}(0,3):=\oint_\Gamma \mathsf{g}(z)\left(GBA_1\Lambda\right)_{j'i}\mathrm{d}z\frac{\partial^3 h_2}{\partial X_{ij}^3}.$$

We first prepare some identities. Using (C.39), observe that

$$\frac{\partial^3 G}{\partial X_{ij}^3}=-6z^{3/2}\left[G\mathcal{L}_{ij}\right]^3 G,$$

which yields that

$$\left(\frac{\partial^3 G}{\partial X_{ij}^3}BD\right)_{j'i}=-6z^{3/2}\big[(G\mathcal{L}_{ij}G\mathcal{L}_{ij}G)_{j'j'}(GBD)_{ii}+(G\mathcal{L}_{ij}G\mathcal{L}_{ij}G)_{j'i}(GBD)_{j'i}\big]. \tag{C.58}$$

Using (C.41) and (C.42), we readily obtain that

$$(G\mathcal{L}_{ij}G\mathcal{L}_{ij}G)_{j'j'}=2(G\mathcal{L}_{ij}G)_{j'i}G_{j'j'},\ (G\mathcal{L}_{ij}G\mathcal{L}_{ij}G)_{j'i}=(G\mathcal{L}_{ij}G\Lambda)_{j'i}G_{ii}+(G\mathcal{L}_{ij}\Lambda G)_{ii}G_{j'i}. \tag{C.59}$$

We summarize the results in the following lemma. Recall (4.16).

Lemma C.7. We have the following estimates:

$$\frac{\kappa_4}{6}\frac{\sqrt{\eta}}{M^{3/2}}\sum_{i,j}\mathsf{P}_{32}(1,2)=\kappa_4\mathsf{V}_2(\mathbf{b},\mathbf{b})\mathcal{Y}^{k-2}+\mathrm{O}_\prec\left(\frac{1}{\sqrt{M\eta}}\right), \tag{C.60}$$
$$\frac{\kappa_4}{6}\frac{\sqrt{\eta}}{M^{3/2}}\sum_{i,j}\mathsf{P}_{32}(2,1)=\mathrm{O}_\prec\left(\frac{1}{M\sqrt{\eta}}\right), \tag{C.61}$$
$$\frac{\kappa_4}{6}\frac{\sqrt{\eta}}{M^{3/2}}\sum_{i,j}\mathsf{P}_{32}(3,0)=\mathrm{O}_\prec\left(\frac{1}{M\sqrt{\eta}}\right), \tag{C.62}$$
$$\frac{\kappa_4}{6}\frac{\sqrt{\eta}}{M^{3/2}}\sum_{i,j}\mathsf{P}_{32}(0,3)=\mathrm{O}_\prec\left(\frac{1}{M\sqrt{\eta}}\right). \tag{C.63}$$

Proof. (1). Justification of (C.60). As before, we first study discussion on some typical terms. Especially, we focus on the following term

$$\frac{\kappa_4\sqrt{\eta}}{2M^{3/2}}\oint_\Gamma \mathsf{g}(z)\frac{\partial^2\mathcal{Z}}{\partial X_{ij}^2}\mathrm{d}z\oint_\Gamma \mathsf{g}(z)\left(\frac{\partial G}{\partial X_{ij}}BA_1\Lambda\right)_{j'i}\mathrm{d}z.$$

We analyze several terms according to (C.37) and (C.21). For notational convenience, we set $D = A_1\Lambda$. We claim that

$$\frac{\kappa_4\sqrt{\eta}}{2M^{3/2}}\oint_\Gamma\oint_\Gamma\sum_{i,j}\mathsf{g}(z_1)\mathsf{g}(z_2)(-\sqrt{z}_1 G_{j'j'}(z_1)(\Lambda G(z_1)BD(z_1))_{ii})z_2\sqrt{M\eta}(\Lambda G(z_2))_{j'j'}(\Lambda G(z_2)BG(z_2))_{ii}$$
$$= -\frac{\kappa_4\eta}{2}\oint_\Gamma\oint_\Gamma \mathsf{g}(z_1)\mathsf{g}(z_2)z_1^{1/2}z_2 m(z_1)m(z_2)\mathcal{J}(z_1,z_2)\mathrm{d}z_1\mathrm{d}z_2 + \mathrm{O}_\prec((M\eta)^{-1/2}), \tag{C.64}$$

where $\mathcal{J}(z_1,z_2)$ is defined as

$$\mathcal{J}(z_1,z_2) := \sum_i(\Sigma_0^{1/2}\Pi_1(z_1)\mathbf{b}\mathbf{b}^*D_1(z_1))_{ii}(\Sigma_0^{1/2}\Pi_1(z_2)\mathbf{b}\mathbf{b}^*\Pi_1(z_2))_{ii},$$

where we recall the definition (C.46).

To see (C.64), by (C.44), we notice that

$$\begin{aligned}
&\frac{\eta}{M}\sum_{i,j}G_{j'j'}(z_1)(\Lambda G(z_1)BD(z_1))_{ii}(\Lambda G(z_2))_{j'j'}(\Lambda G(z_2)BG(z_2))_{ii}\\
&= \frac{\eta}{M}\sum_{i,j}G_{j'j'}(z_1)G_{j'j'}(z_2)(\Sigma_0^{1/2}G_1(z_2)\mathbf{b}\mathbf{b}^*G_1(z_2))_{ii}(\Sigma_0^{1/2}G_1(z_1)\mathbf{b}\mathbf{b}^*D_1(z_1))_{ii}\\
&= \left(\frac{1}{M}\sum_j G_{j'j'}(z_1)G_{j'j'}(z_2)\right)\left(\eta\sum_i[\mathbf{v}_i^*G_1(z_2)\mathbf{b}][\mathbf{b}^*G_1(z_2)\mathbf{f}_i][\mathbf{v}_i^*G_1(z_1)\mathbf{b}][\mathbf{b}^*D_1(z_1)\mathbf{f}_i]\right)\\
&:= \mathcal{L}_1\mathcal{L}_2,
\end{aligned}$$

where we recall (C.55). On one hand, we have from Lemma C.1 that

$$\mathcal{L}_1 = m(z_1)m(z_2) + \mathrm{O}_\prec\left(\frac{1}{\sqrt{M\eta}}\right).$$

On the other hand, by a discussion similar to (C.57), together with Lemma C.1, we obtain that

$$\mathcal{L}_2 - \eta\sum_i[\mathbf{v}_i^*\Pi_1(z_2)\mathbf{b}][\mathbf{b}^*\Pi_1(z_2)\mathbf{f}_i][\mathbf{v}_i^*\Pi_1(z_1)\mathbf{b}][\mathbf{b}^*D_1(z_1)\mathbf{f}_i] = \mathrm{O}_\prec(M^{-1/2}).$$

Consequently, we have that

$$\mathcal{L}_1\mathcal{L}_2 = \eta m(z_1)m(z_2)\sum_i[\mathbf{v}_i^*\Pi_1(z_2)\mathbf{b}][\mathbf{b}^*\Pi_1(z_2)\mathbf{f}_i][\mathbf{v}_i^*\Pi_1(z_1)\mathbf{b}][\mathbf{b}^*D_1(z_1)\mathbf{f}_i] + \mathrm{O}_\prec((M\eta)^{-1/2}).$$

Analogously, by Lemma C.1, using (C.45) and (C.57), we can show that

$$\begin{aligned}&\left|\frac{\eta}{M}\sum_{i,j} G_{j'j'}(z_1)(\Lambda G(z_1)BD(z_1))_{ii}(\Lambda G(z_2))_{ii}(\Lambda G(z_2)BG(z_2))_{j'j'}\right|\\ &\prec \frac{\eta}{M}\frac{M}{M\eta}\sum_i |\mathbf{v}_i^* G_1(z_1)\mathbf{b}||\mathbf{b}^* D_1(z_1)\mathbf{f}_i| \prec \frac{1}{M\sqrt{\eta}},\end{aligned} \tag{C.65}$$

and

$$\begin{aligned}&\left|\frac{\eta}{M}\sum_{i,j}(G\Lambda)_{j'i}(z_1)(G(z_1)BD(z_1))_{j'i}(\Lambda G(z_2))_{j'j'}(\Lambda G(z_2)BG(z_2))_{ii}\right|\\ &\prec \frac{\eta}{M}\frac{M}{M\eta}\sum_i |\mathbf{v}_i^* G_1(z_2)\mathbf{b}||\mathbf{b}^* D_1(z_1)\mathbf{f}_i||\mathbf{b}^* D_1(z_2)\mathbf{f}_i| \prec \frac{1}{M\sqrt{\eta}}.\end{aligned}$$

The rest of the terms can be analyzed since they can all be reduced to the form (C.65). This completes the proof using the above estimates and (C.29).

(2). Justification of (C.61). According to (C.40) and (C.30), we focus on the following term, which is the leading term

$$\begin{aligned}&\left|\frac{\sqrt{\eta}}{M^{3/2}}\sqrt{M\eta}\sum_{i,j}(G(z_1)\Lambda)_{j'i}G(z_1)_{j'j'}(G_1(z_1)\mathbf{b}\mathbf{b}^*D_1(z_1))_{ii}(G(z_2)BG(z_2)\Lambda)_{j'i}\right|\\ &\prec \frac{\eta}{M}\frac{1}{M\eta}\sum_{i,j}|\mathbf{f}_i^* G_1(z_1)\mathbf{b}||\mathbf{b}^* D_1(z_1)\mathbf{f}_i||\mathbf{v}_i^* G_1(z_2)\mathbf{b}| \prec \frac{1}{M\sqrt{\eta}}.\end{aligned}$$

Here in the first step we used Lemma C.1 and in the second step we used a discussion similar to (C.65). The other terms can be analyzed similarly. This completes our proof.

(3). Justification of (C.62). According to (C.58), (C.59), (C.42), and (C.43), we focus our discussion on the following terms, which is the leading term

$$\begin{aligned}&\left|\frac{\sqrt{\eta}}{M^{3/2}}\sum_{i,j} G_{j'j'}(G\Lambda)_{j'i}G_{ii}(G_1\mathbf{b}\mathbf{b}^*D_1)_{ii}\right|\\ &\prec \frac{1}{M}\sum_i |\mathbf{f}_i^* G_1\mathbf{b}||\mathbf{f}_i^* D_1\mathbf{b}| \prec \frac{1}{M\sqrt{\eta}},\end{aligned}$$

where in the first step we used Lemma C.1 and in the second step we used (C.53). The other terms can be studied similarly, and this completes the proof.

(4). Justification of (C.63). According to (C.37), (C.20), (C.41), and (C.42), we have found that it suffices to focus on the following leading term:

$$\left|\frac{\eta}{M}\sum_{i,j}(G(z_1)BA_1(z_1)\Lambda)_{j'i}G_{j'j'}(z_2)(\Lambda G(z_2))_{ii}(\Lambda G(z_2)BG(z_2)\Lambda)_{j'i}\right|$$
$$\prec \frac{\eta}{M^2\eta}\sum_{i,j}|\mathbf{f}_i^*\Pi_1(z_1)\mathbf{b}||\mathbf{b}^*G_1(z_2)\mathbf{f}_i| \prec \frac{1}{M\sqrt{\eta}}.$$

The other terms can be analyzed similarly. This completes our proof. ■

C.4.3 *The error term R_1*

Finally, we control the error term R_1 in the cumulant expansion to complete the verification of (C.28). Recall (C.23). According to Lemma C.4, it suffices to control the following two terms:

$$\mathcal{E}_1 := \sqrt{M\eta}\sum_{i,j}\mathbb{E}\left|X_{ij}^5\mathbf{1}_{\{|X_{ij}|>N^{\epsilon-1/2}\}}\right|\cdot\left\|\frac{\partial^4 w}{\partial X_{ij}^4}\right\|_\infty,\ w=\oint_\Gamma \mathsf{g}(z)h_1\mathrm{d}zh_2,$$

and

$$\mathcal{E}_2 := \sqrt{M\eta}\sum_{i,j}\mathbb{E}|X_{ij}^5|\cdot\sup_{|x|\le N^{\epsilon-1/2}}\left|\frac{\partial^4 w(x)}{\partial X_{ij}^4}\right|.$$

By Lemma C.4, it is easy to see that $R_1 \prec M^{-1/2}$, which follows from the lemma below. Its proof is similar to the discussions in Sections C.4.1 and C.4.2 and we only provide the key points.

Lemma C.8. We have that

$$\mathcal{E}_1,\mathcal{E}_2 \prec M^{-1/2}.$$

Proof. Using an argument similar to the previous subsections on the control of $\partial^k w/\partial X_{ij}^k, 1\le k\le 3$, we can show that

$$\left|\frac{\partial^4 w}{\partial X_{ij}^4}\right| \prec \frac{1}{\sqrt{M\eta}}. \tag{C.66}$$

For $\mathcal{E}_1$, using the assumption (4.5), we find that for any fixed large constant $D>0$,

$$\mathbb{E}\left|X_{ij}^5\mathbf{1}_{\{|X_{ij}|>N^{\epsilon-1/2}\}}\right| \le N^{-D}.$$

Similar arguments hold for $\mathcal{E}_2$ using (4.5) and (C.66). This completes our proof. ■

C.5 The spiked case: CLT for $\widetilde{\mathcal{Y}}$

In this subsection, we briefly discuss how to handle the spiked model and establish the CLT for $\widetilde{\mathcal{Y}}$ as in (4.12). Due to similarity, we focus on explaining the main differences from $\widetilde{Y}$. We will utilize the following identity. It reveals the message that the spiked model can be efficiently reduced to the non-spiked model so that the arguments of Sections C.2–C.4 apply.

Lemma C.9. Recall that $\mathbf{D}=\mathrm{diag}\{d_1,d_2,\cdots,d_r\}$ and $\mathbf{V}_r$ be the collection of the first r eigenvectors. Then we have that

$$\widetilde{G}_1(z)=\Sigma^{-1/2}\Sigma_0^{1/2}\left[G_1(z)-zG_1(z)\mathbf{V}_r(\mathbf{D}^{-1}+1+z\mathbf{V}_r^*G_1(z)\mathbf{V}_r)^{-1}\mathbf{V}_r^*G_1(z)\right]\Sigma_0^{1/2}\Sigma^{-1/2}.$$

Proof. See Lemma C.1 of [33]. ■

According to Lemma C.9, we have that

$$\mathbf{b}^*\widetilde{G}_1(z)\mathbf{b}=\sum_{i=1}^N\frac{\omega_i^2}{1+d_i}\left(\mathbf{v}_i^*G_1(z)\mathbf{v}_i-z\mathbf{v}_i^*G_1(z)\mathbf{V}_r(\mathbf{D}^{-1}+I+z\mathbf{V}_r^*G_1(z)\mathbf{V}_r)^{-1}\mathbf{V}_r^*G_1(z)\mathbf{v}_i\right),$$

where we used the convention that $d_i\equiv 0, i>r$. Denote

$$\Delta(z)=\mathbf{V}_r^*(G_1(z)-\Pi_1(z))\mathbf{V}_r, \tag{C.67}$$

and

$$\mathbf{H}:=(\mathbf{D}^{-1}+I+z\mathbf{V}_r^*G_1(z)\mathbf{V}_r)^{-1},\ \mathbf{L}_1:=(\mathbf{D}^{-1}+I+z\mathbf{V}_r^*\Pi_1(z)\mathbf{V}_r)^{-1}.$$

Then applying a resolvent expansion till the order of two leads to

$$\mathbf{H}=\mathbf{L}_1+\mathbf{L}_1\Delta(z)\mathbf{L}_1+(\mathbf{L}_1\Delta(z))^2\mathbf{H}.$$

We now pause to provide the following control.

Lemma C.10. We have that for some constant $\vartheta>0$

$$\sup_{z\in\Gamma}\|\mathbf{L}_1(z)\|\geq\vartheta.$$

Proof. Note that for $1\leq i\leq r$, we have that $d_i^{-1}+1+z\mathbf{v}_i^*\Pi_1(f(-\widetilde{\sigma}_i^{-1}))\mathbf{v}_i=0$, where $f(\cdot)$ is defined in (4.3). Consequently, according to Assumption 3, we see that for some constant $C>0$,

$$\sup_{z\in\Gamma}\left|d_i^{-1}+1+z\mathbf{v}_i^*\Pi_1(f(z))\mathbf{v}_i\right|=\sup_{z\in\Gamma}\left|\frac{1}{1-\widetilde{\sigma}_i^{-1}\sigma_i}-\frac{1}{1+z\sigma_i}\right|\geq C|\widetilde{\sigma}_i^{-1}-z|\geq\vartheta.$$

This completes our proof. ■

By Lemmas C.1 and C.10, we have that

$$\mathbf{b}^*\widetilde{G}_1(z)\mathbf{b} = \sum_{i=1}^{N} \frac{\omega_i^2}{1+d_i}\left(\mathbf{v}_i^* G_1(z)\mathbf{v}_i - z\mathbf{v}_i^* G_1(z)\mathbf{V}_r\mathbf{L}_1\mathbf{V}_r^* G_1(z)\mathbf{v}_i - z\mathbf{v}_i^*\Pi_1(z)\mathbf{V}_r\mathbf{L}_1\Delta(z)\mathbf{L}_1\mathbf{V}_r^*\Pi_1(z)\mathbf{v}_i\right) + \mathrm{O}_\prec\left(\frac{1}{M\eta}\right).$$

Denote

$$\mathbf{K} := \sum_{i=1}^{N} \frac{\omega_i^2}{1+d_i}\left(\mathbf{v}_i^*\Pi_1(z)\mathbf{v}_i - z\mathbf{v}_i^*\Pi_1(z)\mathbf{V}_r\mathbf{L}_1\mathbf{V}_r^*\Pi_1(z)\mathbf{v}_i\right).$$

Applying Lemma C.1, we have that

$$\mathbf{b}^*\widetilde{G}_1(z)\mathbf{b} - \mathbf{K} = \operatorname{Tr}\left((G_1(z) - \Pi_1(z))\mathbf{A}\right) + \mathrm{O}_\prec\left(\frac{1}{M\eta}\right),$$

where $\mathbf{A}$ is defined as

$$\mathbf{A} := \sum_{i=1}^{N} \frac{\omega_i}{1+d_i}\Big(\mathbf{v}_i\mathbf{v}_i^* - z\mathbf{V}_r\mathbf{L}_1\mathbf{V}_r^*\Pi_1(z)\mathbf{v}_i\mathbf{v}_i^* - z\mathbf{v}_i\mathbf{v}_i^*\Pi_1(z)\mathbf{V}_r\mathbf{L}_1\mathbf{V}_r^* - z\mathbf{V}_r\mathbf{L}_1\mathbf{V}_r^*\Pi_1(z)\mathbf{v}_i\mathbf{v}_i^*\Pi_1(z)\mathbf{V}_r\mathbf{L}_1\mathbf{V}_r^*\Big),$$

where we used the definition (C.67).

To ease our discussion, we denote

$$\mathbf{l}_i := z\mathbf{V}_r\mathbf{L}_1\mathbf{V}_r^*\Pi_1(z)\mathbf{v}_i \tag{C.68}$$

so that we can rewrite

$$\mathbf{A} := \sum_{i=1}^{N} \frac{\omega_i}{1+d_i}\left(\mathbf{v}_i\mathbf{v}_i^* - \mathbf{l}_i\mathbf{v}_i^* - \mathbf{v}_i\mathbf{l}_i^* - z^{-1}\mathbf{l}_i\mathbf{l}_i^*\right).$$

Similar to (C.12), by setting

$$A := \begin{pmatrix} \mathbf{A} & 0 \\ 0 & 0 \end{pmatrix},$$

we find that it suffices to study the distribution of

$$\oint_\Gamma \mathsf{g}(z)\sqrt{M\eta}\,\operatorname{Tr}((G(z) - \Pi(z))A)\mathrm{d}z. \tag{C.69}$$

Compared to (C.12), the only difference is the deterministic part A. The calculations of Sections C.2–C.4 for $\widetilde{\mathcal{Y}}$ still hold here. In what follows, we only explain how to modify the steps. Denote $\widetilde{P}_2$ and $\widetilde{P}_3$ in (C.24) and (C.25) by simply replacing B with A. First, by a discussion similar to (C.31) and (C.32), we can obtain that $\widetilde{P}_2 = -(k-1)\eta\widetilde{\mathcal{L}}\mathcal{Y}^{k-2}$, where $\widetilde{\mathcal{L}}$ is defined similar to (C.32) as follows:

$$\widetilde{\mathcal{L}} := 2\oint_\Gamma\oint_\Gamma \mathsf{g}(z_1)\mathsf{g}(z_2)\,\mathrm{Tr}(\Sigma_0^{1/2}G_1(z_2)\mathbf{A}G_1(z_2)YY^*G_1(z_1)\mathbf{A}(1+m(z_1)\Sigma_0)^{-1}\Sigma_0^{1/2})\mathrm{d}z_1\mathrm{d}z_2.$$

Note that $\widetilde{\mathcal{L}}$ can be controlled using Lemma C.1 as in (C.27) so that we have

$$\widetilde{P}_2 = (k-1)\widetilde{\mathsf{V}}_1\widetilde{\mathcal{Y}}^{k-2} + \mathrm{O}_\prec((N\eta)^{-1/2}),$$

Second, for the high order terms, using an analogous argument, we find that (C.64) holds true by replacing $\mathbf{bb}^*$ with $\mathbf{A}$ and using the fact that $\sum_{i=1}^N \omega_i^2 = 1$ so that as in (C.28) we have

$$\widetilde{P}_3 = (k-1)\kappa_4\widetilde{\mathsf{V}}_2\widetilde{\mathcal{Y}}^{k-2} + \mathrm{O}_\prec((N\eta)^{-1/2}).$$

This completes our proof.

Appendix D. Density and Jacobi Matrix Approximation

In this section, we first discuss a method to compute an approximation of measures of the form (2.16) given a (possibly random) approximation $r(z)$ of

$$\int_{\mathbb{R}} \frac{\mu(\mathrm{d}\lambda)}{\lambda - z}, \quad \mathrm{Im}\, z > 0.$$

We assume that a_j, b_j and c_j are all known, or are well approximated. The approach uses the Chebyshev polynomials of the second kind $(U_k)_{k\geq 0}$ [63], which are the orthogonal polynomials with respect to the semicircle distribution, scaled to $[-1,1]$:

$$\int_{-1}^1 U_k(x)U_j(x)\frac{2\sqrt{1-x^2}}{\pi}\mathrm{d}x = \delta_{jk}.$$

From [73, Lemma 5.6],

$$\int_{-1}^1 \frac{U_k(x)}{x-z}\frac{2\sqrt{1-x^2}}{\pi}\mathrm{d}x = -2\left[z - \sqrt{z-1}\sqrt{z+1}\right]^{k+1} = c_k(z;\mu_{\mathrm{Cheb}}),$$

$$\mu_{\mathrm{Cheb}}(\mathrm{d}x) = \frac{2\sqrt{1-x^2}}{\pi}\mathbf{1}_{[-1,1]}(x)\mathrm{d}x.$$

We then define the mapped polynomials for $a < b$

$$U_k(x;a,b) = U_k(M_{a,b}^{-1}(x)), \quad M_{a,b}(x) = \frac{b-a}{2}x + \frac{b+a}{2}.$$

It is then straightforward to see that

$$\int_a^b \frac{U_k(x;a,b)}{x-z}\sqrt{(b-x)(x-a)}\mathrm{d}x = \frac{\pi(b-a)}{4}c_k(M_{a,b}^{-1}(z);\mu_{\mathrm{Cheb}}).$$

So, given a (small) integer ℓ and unknown coefficients $d_{j,k}$, we can follow the idea of [22] to simply compute $\int \frac{\nu(\mathrm{d}\lambda)}{\lambda - z}$ if ν is of the form (2.16) and

$$h_j(\lambda) = \sum_{k=0}^{\ell-1} d_{j,k} U_k(\lambda; \mathrm{a}_j, \mathrm{b}_j).$$

Let $\mathbf{x}^{(k)} = (x_1^{(k)}, \ldots, x_k^{(k)}) = (x_1, \ldots, x_k)$ be the k zeros of U_k and define the $k \times \ell$ matrix $E_k = (U_{j-1}(x_i))_{\substack{1 \le i \le k \\ 1 \le j \le \ell}}$. This is defined so that

$$h_j(\mathbf{x}^{(k)}) = E_k \begin{bmatrix} d_{j,0} \\ d_{j,1} \\ \vdots \\ d_{j,\ell-1} \end{bmatrix}.$$

For a vector $\mathbf{z} = [z_1, \ldots, z_m]$ of m points in the upper-half plane, define the $m \times \ell$ matrix $C_{\mathbf{z}} = (c_{j-1}(z_i; \mu_{\mathrm{Cheb}}))_{\substack{1 \le i \le m \\ 1 \le j \le \ell}}$.

In the non-spiked case, we seek a solution of the following constrained optimization problem:

$$\operatorname{argmin}_{\mathbf{d}_j : E_k \mathbf{d}_j \ge 0} \left\| \sum_{j=1}^{g+1} \frac{\pi}{4} (\mathrm{b}_j - \mathrm{a}_j) C_{M^{-1}_{\mathrm{a}_j,\mathrm{b}_j}(\mathbf{z})} \mathbf{d}_j - r(\mathbf{z}) \right\|_2 ,$$

where $\mathbf{d}_j = \begin{bmatrix} d_{j,0} & d_{j,1} & \cdots & d_{j,\ell-1} \end{bmatrix}$. If there are spikes c_j, one can approximate the weights w_j using the trapezoidal rule around a small circle with center at c_j. Then the above constrained optimization problem applies to $r(z) - \sum_{j=1}^{p} \frac{w_j}{\mathrm{c}_j - z}$.

Once, the density is approximated, one would like to generate $\mathcal{J}(\mu)$. The simplest way to do this is to use the Gaussian quadrature rule associated to the weight $\sqrt{1-x^2}$, that is, consider the measure

$$\mu_K = \sum_{j=1}^{K} w_j \delta_{x_j^{(K)}},$$

where the weights $\mathbf{w}_K = [w_1, \ldots, w_K]^T$ are chosen so that $\int p(x) \mu_K(\mathrm{d}x) = \int_{-1}^{1} p(x) \frac{2\sqrt{1-x^2}}{\pi} \mathrm{d}x$ whenever p is a polynomial of degree at most $2K - 1$. There are many ways to generate these weights, see [44]. Then define vectors of nodes and weights, respectively, by

$$\mathbf{x} = \begin{bmatrix} M_{\mathrm{a}_1,\mathrm{b}_1}(\mathbf{x}^{(K)}) \\ \vdots \\ M_{\mathrm{a}_{g+1},\mathrm{b}_{g+1}}(\mathbf{x}^{(K)}) \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} \frac{\mathrm{b}_1 - \mathrm{a}_1}{2} (E_K \mathbf{d}_1) \mathbf{w}_K \\ \vdots \\ \frac{\mathrm{b}_{g+1} - \mathrm{a}_{g+1}}{2} (E_K \mathbf{d}_{g+1}) \mathbf{w}_K \end{bmatrix}.$$

If spikes are present, one needs to append $[c_1, \ldots, c_p]$ and $[\omega_1, \ldots, \omega_p]$ onto the end of $\mathbf{x}$ and $\mathbf{W}$, respectively. Now, it follows, in the notation (5.6) that $T(\mathrm{diag}(\mathbf{x}), \sqrt{\mathbf{W}})$, is a good approximation of $\mathcal{J}_K(\mu)$, see [16], for example. Indeed, if we ignore the errors induced by our approximations of each h_j, ω_j, provided $K > K' + \ell/2$ one has that the upper-left $K' \times K'$ block of $T(\mathrm{diag}(\mathbf{x}), \sqrt{\mathbf{W}})$ coincides with that of $\mathcal{J}(\mu)$.

In practice, we generate 100 independent copies of a spiked sample covariance matrix and for each matrix we compute $r(z) = \langle \mathbf{b}, (W - zI)^{-1}\mathbf{b}\rangle$ and take set the points $\mathbf{z}$ to be the union of $M_{a_j,b_j}(\mathbf{u}) + \mathrm{i}/10$ where $\mathbf{u}$ is m equally spaced points on $[-1, 1]$. We take $\ell = 4, m = 200, k = 20$ in our computations. The resulting 100 vectors $\mathbf{d}_j$ are averaged for each j. We do not address the accuracy of this algorithm beyond noting that it suffices to identify the limiting curves in our computations.

Funding

This work is partially supported by National Science Foundation [DMS-2113489 to X.C.D. and DMS-1945652 to T.T.].

Acknowledgments

The authors would like to thank Percy Deift for many useful discussions, Peter Miller for pointing us to references related to Nuttall's theorem, and Deniz Bilman for his superior Tikz abilities. This work was facilitated through the use of advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system at the University of Washington.

References

[1] Aptekarev, A. I. and W. Van Assche. "Scalar and matrix Riemann–Hilbert approach to the strong asymptotics of Padé approximants and complex orthogonal polynomials with varying weight." *J. Approx. Theory* 129, no. 2 (2004): 129–66. https://doi.org/10.1016/j.jat.2004.06.001.

[2] Bai, Z. D., B. Q. Miao, and G. M. Pan. "On asymptotics of eigenvectors of large sample covariance matrix." *Ann. Probab.* 35, no. 4 (2007): 1532–72. https://doi.org/10.1214/009117906000001079.

[3] Bai, Z. and J. W. Silverstein. "Spectral analysis of large dimensional random matrices." *Springer Series in Statistics*, 2nd ed. New York: Springer, 2010.

[4] Baik, J., T. Kriecherbauer, K. T.-R. McLaughlin, and P. D. Miller. *Discrete Orthogonal Polynomials*. Princeton, NJ: Princeton University Press, 2007.

[5] Baik, J., P. Deift, and K. Johansson. "On the distribution of the length of the longest increasing subsequence of random permutations." *J. Amer. Math. Soc.* 12, no. 4 (1999): 1119–78. https://doi.org/10.1090/S0894-0347-99-00307-0.

[6] Bao, Z., K. Schnelli, and X. Yuanyuan. "Central limit theorem for mesoscopic eigenvalue statistics of the free sum of matrices." *Int. Math. Res. Not.* (2020): rnaa210.

[7] Belokolos, E. D., A. I. Bobenko, V. Z. Enol'skii, A. R. Its, and V. B. Matveev. *Algebro-Geometric Approach to Nonlinear Integrable Equations*. Springer, 1994.

[8] Bertola, M., B. Eynard, and J. Harnad. "Differential systems for biorthogonal polynomials appearing in 2-matrix models and the associated Riemann–Hilbert problem." *Comm. Math. Phys.* 243, no. 2 (2003): 193–240. https://doi.org/10.1007/s00220-003-0934-1.

[9] Bertola, M., M. Gekhtman, and J. Szmigielski. "The Cauchy two-matrix model." *Comm. Math. Phys.* 287, no. 3 (2009): 983–1014. https://doi.org/10.1007/s00220-009-0739-y.

[10] Bleher, P. and A. Its. "Semiclassical asymptotics of orthogonal polynomials, Riemann–Hilbert problem, and universality in the matrix model." *Ann. Math.* 150, no. 1 (1999): 185–266. https://doi.org/10.2307/121101.

[11] Bleher, P. and A. B. J. Kuijlaars. "Large n limit of Gaussian random matrices with external source. I." *Comm. Math. Phys.* 252, no. 1–3 (2004): 43–76. https://doi.org/10.1007/s00220-004-1196-2.

[12] Bleher, P. M. *Lectures on Random Matrix Models*. 251–349. New York: Springer, 2011.

[13] Bobenko, A. I. and L. A. Bordag. "Periodic multiphase solutions of the Kadomsev–Petviashvili equation." *J. Phys. A* 22 (1989): 1259–74. https://doi.org/10.1088/0305-4470/22/9/016.

[14] Böttcher, A. and Y. I. Karlovich. *Carleson Curves, Muckenhoupt Weights, and Toeplitz Operators*. Basel: Birkhäuser Basel, 1997.

[15] Breuer, J. and M. Duits. "Universality of mesoscopic fluctuations for orthogonal polynomial ensembles." *Comm. Math. Phys.* 342, no. 2 (2016): 491–531. https://doi.org/10.1007/s00220-015-2514-6.

[16] Brubeck, P. D., Y. Nakatsukasa, and L. N. Trefethen. "Vandermonde with Arnoldi." *SIAM Rev.* 63, no. 2 (2021): 405–15. https://doi.org/10.1137/19M130100X.

[17] Chen, Y. and A. R. Its. "A Riemann–Hilbert approach to the Akhiezer polynomials." *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* 366, no. 1867 (2008): 973–1003. https://doi.org/10.1098/rsta.2007.2058.

[18] Deift, P. *Orthogonal Polynomials and Random Matrices: A Riemann–Hilbert Approach*. Providence, RI: Amer. Math. Soc., 2000.

[19] Deift, P., D. Gioev, T. Kriecherbauer, and M. Vanlessen. "Universality for orthogonal and symplectic Laguerre-type ensembles." *J. Stat. Phys.* 129, no. 5–6 (2007): 949–1053. https://doi.org/10.1007/s10955-007-9325-x.

[20] Deift, P., A. Its, A. Kapaev, and X. Zhou. "On the algebro-geometric integration of the Schlesinger equations." *Commun. Math. Phys.* 203 (1999): 613–33. https://doi.org/10.1007/s002200050037.

[21] Deift, P., T. Kriecherbauer, K. T.-R. McLaughlin, S. Venakides, and X. Zhou. "Uniform asymptotics for polynomials orthogonal with respect to varying exponential weights and applications to universality questions in random matrix theory." *Comm. Pure Appl. Math.* 52, no. 11 (1999): 1335–425. https://doi.org/10.1002/(SICI)1097-0312(199911)52:11\⟨\)1335::AID-CPA1\⟩\)3.0.CO;2-1.

[22] Deift, P., T. Kriecherbauer, and K. T. R. McLaughlin. "New results on the equilibrium measure for logarithmic potentials in the presence of an external field." *J. Approx. Theory* 95, no. 3 (1998): 388–475. https://doi.org/10.1006/jath.1997.3229.

[23] Deift, P., S. D. Miller, and T. Trogdon. "Stopping time signatures for some algorithms in cryptography." (2019): arXiv preprint arXiv:1905.08408.

[24] Deift, P. and X. Zhou. "A steepest descent method for oscillatory Riemann–Hilbert problems. asymptotics for the MKdV equation." *Ann. Math.* 137, no. 2 (1993): 295–368. https://doi.org/10.2307/2946540.

[25] Deift, P. "Riemann–Hilbert methods in the theory of orthogonal polynomials." *Spectral Theory and Mathematical Physics: A Festschrift in Honor of Barry Simon's 60th Birthday*. Proc. Sympos. Pure Math., vol. 76. 715–40. Providence, RI: Amer. Math. Soc., 2007.

[26] Deift, P. and D. Gioev. "Universality at the edge of the spectrum for unitary, orthogonal, and symplectic ensembles of random matrices." *Comm. Pure Appl. Math.* 60 (2007): 867–910. https://doi.org/10.1002/cpa.20164.

[27] Deift, P. and T. Trogdon. "Universality for eigenvalue algorithms on sample covariance matrices." *SIAM J. Numer. Anal.* 55, no. 6 (2017): 2835–62. https://doi.org/10.1137/17M1110900.

[28] Deift, P. and T. Trogdon. "Universality for the toda algorithm to compute the largest eigenvalue of a random matrix." *Commun. Pure Appl. Math.* 71, no. 3 (2018): 505–36. https://doi.org/10.1002/cpa.21715.

[29] Deift, P. and T. Trogdon. "The conjugate gradient algorithm on well-conditioned Wishart matrices is almost deterministic." *Quart. Appl. Math.* 79, no. 1 (2021): 125–61. https://doi.org/10.1090/qam/1574.

[30] Deift, P. A., A. R. Its, and X. Zhou. "A Riemann–Hilbert approach to asymptotic problems arising in the theory of random matrix models, and also in the theory of integrable statistical mechanics." *Ann. Math. (2)* 146, no. 1 (1997): 149–235.

[31] Deift, P. A., G. Menon, S. Olver, and T. Trogdon. "Universality in numerical computations with random data." *Proc. Natl. Acad. Sci.* 111, no. 42 (2014): 14973–8. https://doi.org/10.1073/pnas.1413446111.

[32] Ding, X. "Spiked sample covariance matrices with possibly multiple bulk components." *Random Matrices Theory Appl.* 10, no. 01 (2021): 2150014. https://doi.org/10.1142/S2010326321500143.

[33] Ding, X. and T. Trogdon. "The conjugate gradient algorithm on a general class of spiked covariance matrices." *Q. Appl. Math.* 80, no. 1 (2022): 99–155. https://doi.org/10.1090/qam/1605.

[34] Ding, X. and F. Yang. "A necessary and sufficient condition for edge universality at the largest singular values of covariance matrices." *Ann. Appl. Probab.* 28, no. 3 (2018): 1679–738. https://doi.org/10.1214/17-AAP1341.

[35] Dubrovin, B. A. "Theta functions and non-linear equations." *Russian Math. Surveys* 36 (1981): 11–92. https://doi.org/10.1070/RM1981v036n02ABEH002596.

[36] Duits, M. and A. B. J. Kuijlaars. "Universality in the two-matrix model: a Riemann–Hilbert steepest-descent analysis." *Comm. Pure Appl. Math.* 62 (2009): 1076–153. https://doi.org/10.1002/cpa.20269.

[37] Dumitriu, I. and A. Edelman. "Matrix models for beta ensembles." *J. Math. Phys.* 43 (2002): 5830–47. https://doi.org/10.1063/1.1507823.

[38] Dyson, F. J. "Correlations between eigenvalues of a random matrix." *Comm. Math. Phys.* 19 (1970): 235–50. https://doi.org/10.1007/BF01646824.

[39] Erdoős, L. and H.-T. Yau. *A Dynamical Approach to Random Matrix Theory*. Courant Lecture Notes in Mathematics, vol. 28. Providence, RI: Courant Institute of Mathematical Sciences, New York; American Mathematical Society, 2017.

[40] Fan, Z. and I. Johnstone. "Tracy–Widom at each edge of real covariance and MANOVA estimators." *Ann. Appl. Probab. (in press)* (2021).

[41] Fokas, A. S., A. R. Its, and A. V. Kitaev. "The isomonodromy approach to matrix models in 2D quantum gravity." *Comm. Math. Phys.* 147, no. 2 (1992): 395–430. https://doi.org/10.1007/BF02096594.

[42] Garza, L. E. and F. Marcellán. "Orthogonal polynomials and perturbations on measures supported on the real line and on the unit circle. A matrix perspective." *Expo. Math.* 34, no. 3 (2016): 287–326. https://doi.org/10.1016/j.exmath.2015.12.007.

[43] Gautschi, W. *Orthogonal Polynomials: Applications and Computation*. Oxford University Press, 2004.

[44] Golub, G. H. and J. H. Welsch. "Calculation of Gauss quadrature rules." *Math. Comput.* 23, no. 106 (1969): 221–1, 230. https://doi.org/10.1090/S0025-5718-69-99647-1.

[45] He, Y. and A. Knowles. "Mesoscopic eigenvalue statistics of Wigner matrices." *Ann. Appl. Probab.* 27, no. 3 (2017): 1510–50. https://doi.org/10.1214/16-AAP1237.

[46] Hestenes, M. and E. Steifel. "Method of conjugate gradients for solving linear systems." *J. Research Nat. Bur. Standards* 49 (1952): 409–36. https://doi.org/10.6028/jres.049.044.

[47] Johansson, K. "Random matrices and determinantal processes." *Mathematical Statistical Physics*. 1–55. Amsterdam: Elsevier B. V., 2006.

[48] Johnstone, I. M. "On the distribution of the largest eigenvalue in principal components analysis." *Ann. Stat.* 29, no. 2 (2001): 295–327. https://doi.org/10.1214/aos/1009210544.

[49] Kapaev, A. A. "Riemann–Hilbert problem for bi-orthogonal polynomials." *J. Phys. A* 36, no. 16 (2003): 4629–40. https://doi.org/10.1088/0305-4470/36/16/312.

[50] Khorunzhy, A. M., B. A. Khoruzhenko, and L. A. Pastur. "Asymptotic properties of large random matrices with independent entries." *J. Math. Phys.* 37, no. 10 (1996): 5033–60. https://doi.org/10.1063/1.531589.

[51] Knowles, A. and J. Yin. "Anisotropic local laws for random matrices." *Probab. Theory Relat. Fields* 169, no. 1–2 (2017): 257–352. https://doi.org/10.1007/s00440-016-0730-4.

[52] Krasovsky, I. "Large gap asymptotics for random matrices." *New Trends in Mathematical Physics*, edited by V. Sidoravičius. 413–9. Springer Netherlands, 2009.

[53] Kriecherbauer, T. and K. T.-R. McLaughlin. "Strong asymptotics of polynomials orthogonal with respect to Freud weights." *Internat. Math. Res. Notices* 1999, no. 6 (1999): 299–333. https://doi.org/10.1155/S1073792899000161.

[54] Kuijlaars, A. B. J. and K. T.-R. McLaughlin. "A Riemann–Hilbert problem for biorthogonal polynomials." *J. Comput. Appl. Math.* 178, no. 1–2 (2005): 313–20. https://doi.org/10.1016/j.cam.2004.01.043.

[55] Kuijlaars, A. B. J., K. T.-R. McLaughlin, W. Van Assche, and M. Vanlessen. "The Riemann–Hilbert approach to strong asymptotics for orthogonal polynomials on $[-1,1]$." *Adv. Math.* 188, no. 2 (2004): 337–98. https://doi.org/10.1016/j.aim.2003.08.015.

[56] Kuijlaars, A. B. J. and M. Vanlessen. "Universality for eigenvalue correlations from the modified Jacobi unitary ensemble." *Int. Math. Res. Not.* 2002, no. 30 (2002): 1575–600. https://doi.org/10.1155/S1073792802203116.

[57] Kuijlaars, A. B. J. "Riemann–Hilbert analysis for orthogonal polynomials." *Orthogonal Polynomials and Special Functions (Leuven, 2002)*. Lecture Notes in Math., vol. 1817. 167–210. Berlin: Springer, 2003.

[58] Kuijlaars, A. B. J. and L. Zhang. "Singular values of products of Ginibre random matrices, multiple orthogonal polynomials and hard edge scaling limits." *Comm. Math. Phys.* 332, no. 2 (2014): 759–81. https://doi.org/10.1007/s00220-014-2064-3.

[59] Li, Y., K. Schnelli, and X. Yuanyuan. "Central limit theorem for mesoscopic eigenvalue statistics of deformed Wigner matrices and sample covariance matrices." *Ann. Inst. Henri Poincaré Probab. Stat.* 57, no. 1 (2021): 506–46. https://doi.org/10.1214/20-AIHP1086.

[60] Lytova, A. and L. Pastur. "Central limit theorem for linear eigenvalue statistics of random matrices with independent entries." *Ann. Probab.* 37, no. 5 (2009): 1778–840. https://doi.org/10.1214/09-AOP452.

[61] Mehta, M. L. *Random Matrices*. Elsevier Science, 2004.

[62] Menon, G. and T. Trogdon. "Smoothed analysis for the conjugate gradient algorithm." *SIGMA* 12 (2016): 1–19. https://doi.org/10.3842/SIGMA.2016.109.

[63] Olver, F. W. J., D. W. Lozier, R. F. Boisvert, and C. W. Clark. *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.

[64] Paquette, E. and T. Trogdon. "Universality for the conjugate gradient and MINRES algorithms on sample covariance matrices." *Commun. Pure Appl. Math. (in press)* (2022).

[65] Peherstorfer, F. "Orthogonal polynomials on several intervals: accumulation points of recurrence coefficients and of zeros." *J. Approx. Theory* 163, no. 7 (2011): 814–37. https://doi.org/10.1016/j.jat.2011.03.002.

[66] Pfrang, C. W., P. Deift, and G. Menon. "How long does it take to compute the eigenvalues of a random symmetric matrix? Random matrix theory, interacting particle systems, and integrable systems." *MSRI Publ.* 65 (2014): 411–42.

[67] Sagun, L., T. Trogdon, and Y. LeCun. "Universal halting times in optimization and machine learning." *Q. Appl. Math.* 76, no. 2 (2017): 289–301. https://doi.org/10.1090/qam/1483.

[68] Sankar, A., D. A. Spielman, and S.-H. Teng. "Smoothed analysis of the condition numbers and growth factors of matrices." *SIAM J. Matrix Anal. Appl.* 28, no. 2 (2006): 446–76. https://doi.org/10.1137/S0895479803436202.

[69] Silverstein, J. W. "The smallest eigenvalue of a large dimensional Wishart matrix." *Ann. Probab.* 13, no. 4 (1985): 1364–8. https://doi.org/10.1214/aop/1176992819.

[70] Spielman, D. A. and S.-H. Teng. "Smoothed analysis of algorithms." *J. ACM* 51, no. 3 (2004): 385–463. https://doi.org/10.1145/990308.990310.

[71] Tracy, C. A. and H. Widom. "Level-spacing distributions and the Airy kernel." *Comm. Math. Phys.* 159, no. 1 (1994): 151–74. https://doi.org/10.1007/BF02100489.

[72] Trefethen, L. N. and D. Bau III. *Numerical Linear Algebra*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 1997.

[73] Trogdon, T. and S. Olver. *Riemann–Hilbert Problems, Their Numerical Solution and the Computation of Nonlinear Special Functions*. Philadelphia, PA: SIAM, 2016.

[74] Van Assche, W., J. S. Geronimo, and A. B. J. Kuijlaars. "Riemann–Hilbert problems for multiple orthogonal polynomials." *Special Functions 2000: Current Perspective and Future Directions (Tempe, AZ)*. NATO Sci. Ser. II Math. Phys. Chem., vol. 30. 23–59. Dordrecht: Kluwer Acad. Publ., 2001.

[75] Wang, D. and L. Zhang. "A vector Riemann–Hilbert approach to the Muttalib–Borodin ensembles." (2021): arXiv preprint arXiv 2103.10327.

[76] Yang, F. "Linear spectral statistics of eigenvectors of anisotropic sample covariance matrices." (2020): arXiv preprint arXiv:2005.00999.

[77] Yattselev, M. L. "Nuttall's theorem with analytic weights on algebraic S-contours." *J. Approx. Theory* 190 (2015): 73–90. https://doi.org/10.1016/j.jat.2014.10.015.

[78] Zhedanov, A. "Rational spectral transformations and orthogonal polynomials." *J. Comput. Appl. Math.* 85, no. 1 (1997): 67–86. https://doi.org/10.1016/S0377-0427(97)00130-1.