THE SPECTRUM OF RANDOM KERNEL MATRICES: 
UNIVERSALITY RESULTS FOR ROUGH AND VARYING 
KERNELS 

YEN DO AND VAN VU

Abstract. We consider random matrices whose entries are \( f(X_i^T X_j) \) or \( f(\|X_i - X_j\|^2) \) for iid vectors \( X_i \in \mathbb{R}^p \) with normalized distribution. Assuming that \( f \) is sufficiently smooth and the distribution of \( X_i \)'s is sufficiently nice, El Karoui [17] showed that the spectral distributions of these matrices behave as if \( f \) is linear in the Marchenko–Pastur limit. When \( X_i \)'s are Gaussian vectors, variants of this phenomenon were recently proved for varying kernels, i.e. when \( f \) may depend on \( p \), by Cheng–Singer [13]. Two results are shown in this paper: first it is shown that for a large class of distributions the regularity assumptions on \( f \) in El Karoui’s results can be reduced to minimal; and secondly it is shown that the Gaussian assumptions in Cheng–Singer’s result can be removed, answering a question posed in [13] about the universality of the limiting spectral distribution.

1. Introduction

Let \( X_1, \ldots, X_n \in \mathbb{R}^p \) be iid random vectors with normalization \( \mathbb{E}[X_i] = 0 \) and \( \mathbb{E}\|X_i\|^2 = 1 \), here \( \| \cdot \| \) denotes the Euclidean length in \( \mathbb{R}^p \). For any \( F : \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R} \to \mathbb{R} \) symmetric in the first two variables, consider the random matrix \( A \) with entries

\[
A_{ij} = F(X_i, X_j, p), 
\]

or the variant with zeros on the diagonal

\[
A_{ij} = \begin{cases} 
F(X_i, X_j, p), & i \neq j \\
0, & i = j 
\end{cases}.
\]

Following previous literature [17][13], in this paper these matrices will be referred to as random kernel matrices generated by \( F \) and the distribution of \( X_i \)'s. As described in [17], practical examples of \( F \) are of the form

\[
F(X, Y, p) = f(X^T Y, p) \quad \text{or} \quad f(\|X - Y\|^2, p) ,
\]

More generally, one could have \( F(X, Y, p) = f(g(X, Y), p) \) for some symmetric \( g : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R} \); some Lemmas in this paper are stated in this setting under suitable normalizing assumptions on \( g \) (relative to \( X_i \)'s). For convenience, \( g \) will be referred to as the kernel and \( f \) will be referred to as the envelope that generate \( A \). Examples of envelope functions are \( f(x) = \exp(xa) \), \( f(x) = (1 + x)^a \), where \( a \) is fixed; others can be found in Rasmussen–Williams [30] and Williams–Seeger [39].

\begin{itemize}
\item Date: May 1, 2014.
\item 2000 Mathematics Subject Classification. 15B52.
\item Y. Do is partially supported by NSF grant DMS-1201456.
\item V. Vu is partially supported by NSF grant DMS-0901216.
\end{itemize}
We will be interested in weak-limit of the empirical distribution $\rho_A$ of $A$ when $p, n \to \infty$ such that $p/n \to \gamma \in (0, \infty)$, a fixed constant. Recall that

$$\rho_A(x) := n^{-1} \sum_{i=1}^{n} \delta_{\lambda_i}(x),$$

here $\lambda_1, \ldots, \lambda_n$ are eigenvalues of $A$ and $\delta_\lambda$ denotes the counting measure at $\lambda$. This research direction has been investigated recently by El Karoui [17] and Cheng–Singer [13], motivated by studies from machine learning and statistical analysis. In [17], it was assumed that $A$ is generated by either the inner-product or the distance kernels, and with $p$-independent envelope functions (which is the natural setting relative to the above normalization of $X_i$). It was shown in [17] that for $f$ sufficiently smooth the limiting behavior of $\rho_A$ depends only on a linear component of $f$. It turns out that a variant of this phenomenon continues to hold even if $f$ depends on $p$: for $g(X, Y) = X^T Y$ this was proved for Gaussian random vectors in a recent result of Cheng and Singer [13]. See also Bordenave [9] for a related recent work in this direction that appeared after an initial circulation of a first draft of this paper.

The goal of this paper is to extend the results in [17, 13] to more general settings. In particular, Theorem 3 will (positively) answer a question by Cheng and Singer [13] regarding the universality of the limiting spectral distribution of Cheng–Singer’s models.

We would like to point out that El Karoui [18] also considered a related model where the entries $g(X_i^T X_j)$ are affected by random noise before the envelope $f$ is applied outside; the reader is referred to the beautiful work [18] for further details. There is also a vast amount of literature concerning limiting behaviors of $\rho_A$ when $p$ is low or fixed, the interested reader is referred to [26, 36, 39, 23, 6, 8, 22] and references there-in.

For clarity, below the discussion of previous and new results and are divided into two sections.

1.1. The $p$-independent setting. In this section, the setting when $F$ is independent of $p$ (relative to the above normalization of $X_i$) will be discussed. In other words, only the settings when $F(X, Y, p) = f(X^T Y)$ or $f(\|X - Y\|^2)$ (for some $p$-independent envelope function $f$) will be considered. Since the vectors $X_i$ are normalized, this is the natural setting for $f$.

1.1.1. The inner product kernel. Let $F(X, Y, p) = f(X^T Y)$. When the limiting spectral distribution for the model (1.1) of $A$ is known, standard arguments may be used to deduce the limiting spectral distribution of (1.2) (see e.g. [5] or [17], see also Lemma 2 of the current paper), and vice versa. Below the model (1.1) will be assumed unless otherwise stated.

For linear envelope functions, it is well-known that if the distribution of the vectors $X_i$‘s satisfies certain martingale/concentration properties then $\rho_A$ converges weakly to some form of the Marčenko–Pastur (MP) distribution, whose density is given by

$$\rho_{MP, \gamma}(x) := (1 - \gamma)1_{\{\gamma < 1\}}\delta_0(x) + \frac{\gamma}{2\pi x} \sqrt{(b-x)(x-a)}1_{[a,b]}(x),$$

$$a := (1 - \frac{1}{\sqrt{\gamma}})^2 \quad \text{and} \quad b := (1 + \frac{1}{\sqrt{\gamma}})^2.$$
This convergence was first established by Marčenko–Pastur [25] (see also Wachter [38]) when the entries of each vector $X_i$ are iid. Various authors have then extended this result to more general settings, see e.g. Auburn [3], Yin and Krishnaiah [40], Silverstein [31], Götze and Tikhomirov [19, 20], El Karoui [15, 17], Adamczak [1], Pajor and Pastur [28], Bordenave et al. [7], Chafai [10], Chatterjee et al. [12], and O’Rourke [27]. In particular, the result holds for $X_i$’s drawn independently from isotrophic log-concave distributions, this is a result of Pajor and Pastur [28]. Extensions to settings with some martingale-type assumptions were carried out in [19, 20, 1], and extensions to settings with some concentration conditions on the distributions of $X_i$’s were done in [15]. See also [27, 7, 12, 10] for other generalizations.

For nonlinear envelope functions with sufficient smoothness, it was shown by El Karoui [17] that if the distribution of $X_i$’s is sufficiently nice then $A$ has the same the limiting spectral distribution as

$$B = [f(1) - f(0) - f'(0)]I_n + f'(0)(X_i^T X_j)_{i,j}.$$  

Here and in the rest of the paper, $I_n$ is the identity $n \times n$ matrix. Let $\alpha := f(1) - f(0) - f'(0)$. Using the linear theory, it follows that

$$\rho_A(x) \xrightarrow{w} \begin{cases} \frac{1}{f'(0)}\rho_{MP,\gamma}(\frac{x-\alpha}{f'(0)}), & \text{if } f'(0) \neq 0; \\ \delta_\alpha, & \text{if } f'(0) = 0. \end{cases}$$

In El Karoui [17], the convergence (1.3) was considered in two different settings:

(i) The iid setting with $K$ moment bounds: Assume that

$$\mathbb{E}|\sqrt{p}X_{ij}|^K = O(1).$$

In this setting, it was shown in [17] that (1.3) holds if: $K > 4$ and $f$ is $C^3$ near 0 and $C^1$ near 1.

(ii) The high concentration setting with parameter $c(p)$: Assume that

$$P(|F(X_i) - m_F| > t) \leq C \exp(-c(p)t^b), \quad \forall t > 0,$$

where $m_F$ denotes a median of $F(X_i)$.

In this setting, it was shown in [17] that (1.3) holds if two conditions hold:

- $f$ is $C^2$ near 0 and $C^1$ near 1, and
- $c(p) \geq Cp^{b/4}$ for some absolute constant $C > 0$. (For simplicity we’ll write $c(p) \geq O(p^{b/4}).$)

In the iid setting (1.4), it was shown in [17] that a stronger convergence in spectral norm holds. In particular, it follows that some information about the largest eigenvalue of $B$ could be transferred to $A$. The interested reader is referred to [17, 16] and the references there-in for related literature.

The estimate (1.5) is satisfied for a large class of distributions, including:

(a) $X_i$’s are Gaussian vectors (which is clearly a special case of (1.4));

(b) $X_i$’s are sampled from the unit sphere.

(c) $X_i$’s are sampled from a distribution with log-concave density $e^{-U(x)}$ such that $\text{Hess } U(x) - c(p)I_d$ is positive definite. (In this case $b = 2$, see e.g. [24].)
Other examples can be found in [17, 15] and [24].

In the special cases (a,b) above, El Karoui’s results were improved recently by Cheng–Singer [13], where the authors showed that similar results hold for the variant [1.2] under the weaker assumption that \( f \) is \( C^1 \) near 0.

An initial examination of Karoui’s results reveals that one only needs differentiability of \( f \) at 0 to formulate the above linear component \( B \) of \( A \). On closer looks, perhaps continuity of \( f \) at 1 is also required, since the diagonal entries of the covariance matrix of \( X_i \) are converging to 1 in the large \( n \) large \( p \) limit; except for the zero-diagonal model [1.2].

In the first result of this paper, it will be shown that under these minimal regularity assumptions on \( f \) the nonlinear-to-linear results of [17] can still be proved for a large class of distributions. Similar settings for distribution of \( X_i \) will be considered:

- The iid setting [1.4] with \( K > 4 \) moment bounds.
- The high concentration setting [1.5] with \( c(p) \geq O(p^{b/2}) \).

While our assumption \( c(p) \geq O(p^{b/2}) \) is stronger than El Karoui’s assumption, it is in fact satisfied by a fairly large class of interesting distributions (see [24] and also [17] for many examples); also a recent work of Guédon–Milman [21] (cf. [29]) indicates that such concentration assumption may be true for the isotropic log-concave setting (see also the discussion following Conjecture 4 for details). On the other hand, Theorem 1 requires only minimal regularity assumptions on \( f \); one might view this as a trade-off between regularity of \( f \) and concentration assumptions on the population distribution.

**Theorem 1.** Assume the iid setting [1.4] with \( K > 4 \), or the high concentration setting [1.5] with \( c(p) \geq O(p^{b/2}) \).

(i) Let \( f \) be differentiable at 0 and continuous at 1. Let \( A \) be defined by [1.1] with \( F(X, Y, p) = f(X^TY) \). Then \( A \) has the same limiting spectral distribution as

\[
B = [f(1) - f(0) - f'(0)]I_n + f'(0)(X_i^TX_j)_{i,j}.
\]

(ii) Let \( f \) be differentiable at 0. Let \( A \) be defined by [1.2] with \( F(X, Y, p) = f(X^TY) \). Then \( A \) has the same limiting spectral distribution as

\[
B = [-f(0) - f'(0)]I_n + f'(0)(X_i^TX_j)_{i,j}.
\]

Theorem 1 will be shown in Section 3.

1.1.2. The distance kernel. Let \( F(X, Y, p) = f(\|X - Y\|^2) \). This model has recently attracted the attention of some authors (see e.g. [17, 22, 8]), motivated by connections to machine learning theory and physics. In this model, it is clear that the two settings [1.1] and [1.2] are equivalent up to a shift (by \( f(0) \)) of the limiting spectral distribution. Below it will be assumed that \( A \) is defined using [1.1].

When \( (X_i) \) are Bernoulli or sampled from the unit sphere, the distance kernel model follows from the inner product model, but it is not hard to find interesting examples (such as Gaussian or log-concave) where a naive adaptation of this reduction does not apply. This however suggests that \( A \) should have the same limiting spectral distribution as

\[
B = [f(0) - f(2) + 2f'(2)]I_n - 2f'(2)(X_i^TX_j)_{i,j}.
\]
when $f$ is sufficiently smooth and the distribution of $X_i$ is sufficiently concentrated. This was shown in El Karoui [17], where the author assumed essentially the same settings for the $X_i$’s as in the last section:

- In the iid setting (1.4) with $K > 5$ moment bounds, this was shown for $f$ being $C^3$ near 2.
- In the high concentration setting (1.5) with $c(p) > O(p^{b/2})$ this was shown for $f$ being $C^2$ near 2 and $C^1$ near 0.

In the iid setting, it was furthermore shown in [17] that a stronger convergence holds in spectral norm, which may lead to more information about the distribution of the largest eigenvalue of $A$. (The interested reader is referred to to [17, 15] and references there-in for related literature.) As remarked earlier, the limiting spectral distribution of $B$ may be computed explicitly using Marchenko–Pastur theory.

It is clear that one only requires differentiability of $f$ at 2 to write down the above linear component $B$ of $A$. Theorem 2 below shows that for a large class of distributions of $X_i$, El Karoui’s nonlinear-to-linear results for distance random matrices can be proved for $f$ assuming only this differentiability.

**Theorem 2.** Assume the iid setting (1.4) with $K > 4$, or the high concentration setting (1.5) with $c(p) ≥ O(p^{b/2})$.

Let $A$ be defined using (1.1) with $F(X, Y, p) = f(\|X - Y\|^2)$ where $f$ is differentiable at 2. Then $A$ has the same limiting spectral distribution as

$$B = [(f(0) - f(2) + 2f'(2))I_n - 2f''(2)(X^T X)]_{i,j}.$$ 

Besides regularity improvement, in the iid setting Theorem 2 requires less moment bounds than El Karoui [17]. As discussed before in the paragraph leading to the statement of Theorem 1, our assumption on $c(p)$ is stronger than that in [17], but is satisfied by a large class of interesting distributions, see [24] and [17] for many examples. A recent work of Guédon–Milman [21] (cf. Paouris [29]) indicates that such concentration inequality may hold in the isotropic log-concave setting, see Conjecture 2 for details.

Theorem 2 will be proved in Section 3.

**1.1.3. Conjectures.** We conjecture that similar results hold in the isotropic log-concave case.

**Conjecture 1.** The results of Theorem 1 hold if $X_i$’s are iid random vectors from a normalized isotropic log-concave distribution.

**Conjecture 2.** The results of Theorem 2 hold if $X_i$’s are iid random vectors from a normalized isotropic log-concave distribution.

Recall the following result of Guédon–Milman [21] (cf. Paouris [29]):

**Proposition 1.** [21] Let $X$ be a (normalized) random vector sampled from an isotropic measure in $\mathbb{R}^p$ with log-concave density, then for any $t ≥ 0$

$$\mathbb{P}(\|X\| - 1 ≥ t) ≤ C \exp(-ct^{1/2} \min(t, t^3))$$

for some absolute constants $c, C ∈ (0, ∞)$.

It follows from the proofs of Theorem 1 and Theorem 2 that both conjectures hold if the estimate of Proposition 1 could be improved to $O(\exp(-ct^{1/2}t))$. Mainly, in these proofs the concentration bound (1.5) is only needed for the Lipschitz function $f$ at 2.
$\|X\|$, here $X$ is either $X_i$ or $X_i + X_j$ (which also have an isotropic log-concave distribution). This leads to the question of whether one can remove the term $t^3$ in the Guédon-Milman concentration result.

See also Bordenave \[8\] for a recent work that was motivated by Conjecture \[2\].

1.1.4. Possible extensions of Theorem 1 and Theorem 2. Theorem 1 and Theorem 2 could be considered as extensions of Theorem 2.3 and Theorem 2.4 of El Karoui in \[17\] when the envelope/kernel functions are rougher. On the other hand, in \[17\] El Karoui considered a more general setting when the covariance matrix $\Sigma$ of $X_i$ is less restrictive (in the current paper we assume $\Sigma = Id$ the identity operator in $\mathbb{R}^p$, which is the simplest but also most natural setting). More precisely, in \[17\] $\Sigma$ is allowed to depend on $p$, but is still positive definite and converges in some fashion to identity in the limit $n, p \to \infty$. We anticipate that the proof of Theorem 1 and Theorem 2 in this paper has a natural extension that could lead to an extension of these Theorems to settings similar to those considered in Theorem 2.3 and 2.4 of \[17\], however these extensions are not explored in the current paper.

1.2. The $p$-dependent setting. In this section, a less classical setting recently investigated in \[13\] will be discussed. Here, the envelope function $f$ is allowed to depend on $p$ and may have very little regularity in $x$. In other words, $f$ may be varying with $p$. Examples of such situations and their motivations are presented in \[13\]. In this paper, only the inner-product kernel $g(X, Y) = X^T Y$ will be considered, and a similar investigation for the distance kernel is left for a further study. Furthermore, following \[13\], only the non-diagonal model \[12\] will be considered, and analogous results for the diagonal model \[13\] may be obtained using a diagonal perturbation argument.

In this section, $X_1, X_2, \ldots, X_n$ will be iid random vectors in $\mathbb{R}^p$ whose coordinates are independent copies of a random variable $Z$ with mean 0 and variance $1/p$, such that for all $K > 0$ there is a constant $C_K$ depending on $K$ such that

$$E|Z|^K \leq C_K p^{-K/2}.$$  

While \[16\] requires all $K > 0$, this assumption may be improved if there are better bounds on the growth of a scaled version of $f$ as $p \to \infty$. For details, see the remark after the statement of Theorem 3.

Below, some standard facts about orthogonal polynomials will be recalled, for a standard reference see e.g. \[32\] or \[2\]. Given a nonnegative measure $\mu$ on $\mathbb{R}$ and $k = 0, 1, 2, \ldots$, the $k^{th}$ orthogonal polynomial $p_k(x)$ with respect to $\mu$ is a polynomial of degree $k$ with positive leading coefficient, such that

$$\int_\mathbb{R} p_k(x)p_m(x)d\mu(x) = \begin{cases} 0, & \text{if } m \neq k; \\ 1, & \text{if } m = k. \end{cases}$$

For any function $h \in L^2(d\mu)$ that belongs to the span of $\{p_n, n \geq 0\}$, one has the formal series

$$\sum_{k=0}^{\infty} a_k p_k(x),$$

where $a_k := \langle h, p_k \rangle_\mu = \int_\mathbb{R} h(x)p_k(x)d\mu(x)$,

and if the series converges to $h$ in $L^2(\mu)$ then the Plancherel equality holds: $\|h\|_{L^2(d\mu)}^2 = \sum_{k \geq 0} |a_k|^2$. In that case, since $p_0(x) \equiv 1$ for probability distributions, it follows
that if \( \mu \) is the distribution of some random variable \( \xi \) then
\[
\text{Var}[h(\xi)] = \sum_{k \geq 1} |a_k|^2 .
\]

For each \( p \) let \( \xi_p = \sqrt{p}X^TY \) where \( X \) and \( Y \) are two iid copies of (any) vector \( X_i \). Clearly \( \xi_p \) has mean 0 and variance 1. Let \( p_{k,p}(x) \), \( k \geq 0 \), be the orthogonal polynomials with respect to the probability distribution \( \mu_p \) of \( \xi_p \).

Below we state the conditions that will be assumed on the envelope function \( f(x, p) \) for the next result, Theorem 3. These conditions were first formulated in [13] in an equivalent form. For any \( f(x, p) \) let \( k(x, p) := \sqrt{pf(x/\sqrt{p}, p)} \) and consider the expansion
\[
k(x, p) \sim \sum_{k \geq 0} a_{k,p}p_{k,p}(x) .
\]

In this paper, \( f \) is said to be admissible with respect to the generating distribution of \( X_i \) if the following three conditions hold.

(i) (Uniform convergence) The orthogonal polynomial series of \( k(x, p) \) converges to \( k(x, p) \) in \( L^2(\mu_p) \) uniformly over \( p \) large. In other words, for any \( \epsilon > 0 \) there exists \( L = L(\epsilon) \) such that the following holds for \( p \) large
\[
\|k - \sum_{i \leq L} a_{i,p}p_{i,p}\|_{L^2(\mu_p)}^2 = \sum_{i > L} |a_{i,p}|^2 \leq \epsilon .
\]

(ii) (Normalization) There exists \( \nu \in [0, \infty) \) such that
\[
\lim_{p \to \infty} \sum_{i=1}^{\infty} |a_{i,p}|^2 = \nu .
\]

(iii) (Scaling) There exists \( a \in [0, \infty) \) such that
\[
\lim_{p \to \infty} a_{1,p} = a .
\]

It is clear that (1.8) and (1.9) together imply the condition \( a^2 \leq \nu \), which shall be assumed throughout. It is worth pointing out that the set of orthogonal polynomials with respect to a probability measure \( \mu \) does not always form a complete basis in \( L^2(\mu) \), this however holds for a fairly large class of probability measures, including those with sub-exponential tails (i.e. \( \mathbb{P}(|\xi| > x) = O(e^{-cx}) \) for some \( c > 0 \), see e.g. [2, Theorem 6.5.2]). In particular this completeness holds if the measure is compactly supported. It follows that the convergence of the orthogonal expansion in \( L^2 \) holds automatically if \( X_i \)'s are Gaussian or bounded (note that this gives convergence of the expansion for each \( p \), and condition (1.7) is about the uniformity of the convergence). In the general case when completeness of the orthogonal polynomials is not guaranteed, the condition (1.7) has to be checked carefully (for both the convergence of the expansion for each \( p \), and the uniformity of the convergence over \( p \) large).

For convenience of the reader, the definition of the Stieltjes transform \( m(z) \) of a measure \( \mu \) is recalled below:
\[
m(z) = \int_{\mathbb{R}} \frac{d\mu(x)}{x - z} , \quad \text{Im}(z) > 0 .
\]
Theorem 3. Assume that $f$ is admissible with respect to $X_i$'s which satisfy (1.6). Let $A$ be generated using (1.2) using $F(X,Y,p) = f(X^TY,p)$. Then, the empirical distribution of $A$ converges weakly to a probability distribution whose Stieltjes transform $m(z)$ satisfies:

$$(1.10) \quad -\frac{1}{m(z)} = z + a\left(1 - \frac{1}{1 + \frac{2}{\gamma}m(z)}\right) + \frac{\nu - a^2}{\gamma}m(z).$$

Remark 1. As pointed out in [13], this limiting spectral distribution is no longer MP when $\nu \neq a^2$. Unique solvability of (1.10) was proved in [13] using elementary arguments.

The special case of Theorem 3 for Gaussian random vectors was proved by Cheng–Singer in [13]. Theorem 3 (positively) answered the question of Cheng and Singer in [13] about the validity of their result for more general distribution (such as Bernoulli).

While it is assumed in Theorem 3 that (1.6) holds for all $K > 0$, the method of proof can be easily refined to lessen this assumption when more information about $f$ is given. More precisely, if there is an upper bound $L$ on the cut-off degree $L(\epsilon)$ in (1.7) (independent of $\epsilon > 0$) then we only need to have (1.6) up to $O_L(1)$: the degree $L$ will enter the proof in Lemma 9 and Lemma 8 and will eventually dictate the number of required moment bounds on entries of $X_i$'s. For instance, when $f$ is independent of $p$ and has a non-vanishing derivative at $x = 0$ it can be verified that $\nu = a^2 = f'(0)^2$, and one can see from the proof of Theorem 1 that one could take $L = 1$ in (1.7). Eventually, with some refinements (tailored specifically for Theorem 1), this leads to the requirement $K > 4$ in Theorem 1.

Acknowledgement. We would like to thank X. Cheng and A. Singer for bringing this interesting subject to our attention and many useful conversations. We would like to thank the referees for corrections and suggestions which have lead to improvement of the quality of the paper.

2. The general ideas

Let $m_A(z)$ denote the Stieltjes transform of the empirical spectral distribution $\rho_A$ of $A$: in the following $m_A$ will be refered to as the Stieltjes transform of $A$. Explicitly,

$$m_A(z) = \frac{1}{n} \text{Tr}[(A - z)^{-1}], \quad \text{Im}(z) > 0.$$ 

By standard reductions (see e.g. [5]), it suffices to show that $m_A(z)$ converges to the Stieltjes transform of the desired limiting spectral distribution (which is always a probability distribution in the current paper) for $\text{Im}(z) > 0$. For instance, in the setting of Theorem 3 it will be shown that $m_A$ converges to the solution of (1.10).

The main idea for showing the desired convergence of $m_A$ is to compare $A$ with a suitably chosen random matrix whose Stieltjes transform already has the desired convergence. In fact, due to a result from [13] which asserts that

$$(2.1) \quad \lim_{n \to \infty} |m_A(z) - \mathbb{E}m_A(z)| = 0 \quad \text{a.s.},$$

it suffices to compare expected values of Stieltjes transforms in question. To keep the paper self-contained, a short proof of (2.1) will be included in section 3.2 (see Lemma 3).
The proof for the results in the $p$-independent case will be presented in the next section. Following El Karoui [17], $A$ will be compared with a linear approximation of $A$, obtained by replacing the envelope function $f$ with the linear part of its Taylor expansion at a suitable point. The main idea which allows us to improve the regularity assumptions on $f$ in El Karoui and Cheng–Singer’s results is a simple transference principle, see Lemma 1 and also its companion Lemma 2 in Section 3 for details.

The proof of the results in the $p$-dependent setting will use a series of comparisons. In order to carry out the analysis of the main comparison, the Lindeberg swapping method will be used, following ideas from [11, 34]. This method has recently proved useful in various studies of random matrices, especially for the local statistics (see [35] for a survey). One of the main difficulties in implementing the Lindeberg method in the current setting is the lack of regularity of the envelope function. To overcome this difficulty, the uniform convergence condition (3.1) will be used, as this condition allows for approximation of $f$ with polynomials (which are very smooth). The proof of Theorem 3 will be presented in Section 4.1 and Section 4.2.

In the rest of the paper, without loss of generality it will be assumed that $\text{Im}(z) > 0$. All implicit constants in the paper may depend on $z$. All asymptotics notations are used under the assumption that $p, n \to \infty$.

3. THE $p$-INDEPENDENT SETTING

In this section, we prove Theorem 1 and Theorem 2.

Let $A$ be defined by (1.2) using $f(g(X, Y))$ and let $A$ be defined by (1.2) using $F(X, Y, p) = g(X, Y)$. The following transference principle asserts that one can deduce the limiting spectral density for $A$ from $A$ as long as:

(i) $f$ is differentiable at the mean value of $g(X_i, X_j)$; and

(ii) the entries of $A$ (hence the kernel $g$) satisfy a fairly general concentration condition (relative to $X_i$’s).

Lemma 1 (Transference principle). Assume that $\rho_A$ converges weakly to a probability distribution. Let $a = \text{E}g(X_i, X_j)$ and let $f$ be differentiable at $x = a$. Assume

$$\text{Var}[g(X_i, X_j)] = O(1/p),$$

and assume that for any fixed $\delta > 0$ it holds that

$$P(\max_{i \neq j} |g(X_i, X_j) - a| > \delta) = o(1).$$

Then $A$ has the same limiting spectral distribution as

$$B = (af'(a) - f(a))I_n + f'(a)A.$$ 

Remarks: While different pairs $(f, g)$ may generate the same $F = f \circ g$, the two constraints (3.1) and (3.2) impose a strong normalization on $g$. Also, in Lemma 1 the spectral distribution of $A$ is not required to be Marchenko–Pastur.

The following simple result will also be used, which says that under an assumption on concentration of $g(X_i, X_i)$, the models (1.1) and (1.2) are equivalent.

Lemma 2. Let $A_1$ and $A_2$ be defined by (1.1) and (1.2) respectively using $F(X, Y, p) = f(g(X, Y))$. Assume that $f$ is continuous at $b := \text{E}g(X_i, X_i)$. Assume that for any
\( \delta > 0 \) it holds that
\[
P( \max_{1 \leq i \leq n} |g(X_i, X_i) - b| > \delta) = o(1) \ .
\] (3.3)

Then in the large \( n \) large \( p \) limit it holds that
\[
|m_{A_1}(z + f(b)) - m_{A_1}(z)| = o(1) \ a.s.
\]

Remark 2. If \( g(X_i, X_i) \) is a constant then the above result is trivial, in which case continuity of \( f \) at \( b \) is not needed.

Using Lemma 2 the main argument is reduced to the non-diagonal model (1.2), where the transference principle could be used. Lemma 2 could be viewed as a companion of Lemma 1.

Below, Theorem 1 and Theorem 2 are deduced from the above two Lemmas. Proofs of Lemma 1 and Lemma 2 are presented in Section 3.3.

3.1. Proof of Theorem 1

**The iid case:** Assume that the entries of \( X_i \) are iid with \( K > 4 \) moment bounds.

**Step 1:** We first reduce the Theorem to the model (1.2) of \( A \). By Lemma 2 it suffices to show that: for any fixed \( \delta > 0 \) (i.e. independent of \( n, p \)) and any \( i \) we have
\[
P(\|X_i\|^2 - 1 > \delta) = o(1/p) \ .
\] (3.4)

We note that a more quantitative estimate was proved in [17] using more careful arguments, on the other hand for (3.4) the following simplified argument suffices. Fix \( \delta > 0 \) and let \( X = (x_1, \ldots, x_p) \) be an independent copy of \( X_i \)'s. Let \( M := p^3 \) for \( 0 > \beta > \frac{2}{K} - \frac{1}{2} \). Let \( E = \{\max_i |x_i| > M\} \), clearly
\[
P(E) \leq C \sum_j p^{-\beta K} E|x_j|^K = o(1/p) \ .
\]

Let \( \tilde{X} = (x_j 1_{|x_j| \leq M})_j \). On \( E^c \) clearly \( \tilde{X} = X \). Thus, it suffices to show that
\[
P(\|\tilde{X}\|^2 - 1 > \delta) = o(1/p^2) 
\] (3.5)

Let \( \tilde{1} = (1, \ldots, 1) \in \mathbb{R}^p \). Let \( \mu \) and \( \sigma^2 \) be the mean and variance of \( x_i 1_{|x_i| \leq M} \). It is not hard to see that
\[
\mu = o\left(\frac{1}{\sqrt{p}}\right) \quad \text{and} \quad \sigma^2 = \frac{1}{p} + o\left(\frac{1}{p}\right) 
\]

For \( p \) sufficiently large, it follows that
\[
P(\|\tilde{X}\|^2 - 1 > \delta) \leq P(\|\tilde{X} - \mu \tilde{1}\|^2 - p\sigma^2 < \delta/2) \ .
\] (3.6)

Write \( \|\tilde{X} - \mu \tilde{1}\|^2 - p\sigma^2 = \sum_{j=1}^p [(\tilde{x}_j - \mu)^2 - \sigma^2] \) sum of iid random variables, each has mean 0 and is bounded above by \( O(M^2) = O(p^{2\beta}) \) and has the following variance bound:
\[
N := \text{Var}[(\tilde{x}_j - \mu)^2 - \sigma^2] = \mathbb{E}[(\tilde{x}_j - \mu)^4] - \sigma^4 = O(p^{-2}) 
\]

By Chernoff’s inequality (see e.g. [33]), for \( C_1, C_2 \) absolute positive constants it holds that
\[
P(\|\tilde{X} - \mu \tilde{1}\|^2 - p\sigma^2 > \delta/2) \leq
\]
\[
C_1 \max\left(\exp\left(-\frac{C_3\delta^2}{pnN}\right), \exp\left(-\frac{C_2\delta}{M^2}\right)\right) = o(p^{-1}) \ .
\] (3.7)
Collecting inequalities \((3.6)\) and \((3.7)\), the desired estimate \((3.5)\) follows.

**Step 2:** Thanks to Step 1, it remains to show the theorem for \(A\) given by \((1.2)\). Let \(\delta > 0\) be fixed. Using Lemma \(1\) it suffices to show that

\[
E[|X_i^T X_j|^K] = O(p^{-K/2}) , \quad \text{for some } K > 4.
\]

Write \(X_i = (x_{i1}, \ldots, x_{ip})\) and \(X_j = (x_{j1}, \ldots, x_{jp})\). Then \(x_{ik}\) and \(x_{jm}\) are independent with mean 0 and variance \(1/p\) for any \(1 \leq k, m \leq p\). By the inverse Khintchine inequality (i.e. the Marcinkiewicz–Zygmund inequality) it holds that

\[
\mathbb{E}|X_i^T X_j|^K \leq C_K \mathbb{E} \left( \sum_{m=1}^{p} |x_{im} x_{jm}|^2 \right)^{K/2}
\]

\[
\leq C_K p^{K/2-1} \sum_{m=1}^{p} \mathbb{E}|x_{im}|^K \mathbb{E}|x_{jm}|^K \quad (\text{Hölder, then independence})
\]

\[
= O(p^{-K/2}) \quad \text{(using given moment bounds)}.
\]

The high concentration case: As in the iid case, it suffices to show \((3.5)\) and \((3.8)\). As can be seen below, in the proof it is enough to assume \((1.5)\) for the \(1\)-Lipschitz functions of the form \(f(X) = \|X + c\|\), \(c \in \mathbb{R}^p\) constant vectors. See also the discussion after the statements of Conjecture \(1\) and Conjecture \(2\).

**Proof of \((3.4)\):** Let \(F(Y) = \|Y\|\) the Euclidean length of \(Y \in \mathbb{R}^p\), clearly \(F\) is 1-Lipschitz. Fix \(i\) and \(\delta > 0\), without loss of generality assume \(\delta < 1/2\).

We first show that uniformly over \(r > 0\) it holds that

\[
P(\|X_i\| - 1 > r) = O(e^{-c(p)r^b/C_b}) ,
\]

here and below \(C_b\) will denote absolute constants that could depend on \(b\). Let \(a = \mathbb{E}\|X_i\|\). It suffices to show that \(|a - 1| = O(c(p)^{-1/b})\). By \((1.5)\), it holds that

\[
|a - m_F| \leq \mathbb{E}\|X_i\| - m_F
\]

\[
= O(\int_0^\infty e^{-c(p)r^b} \, dr) = O(c(p)^{-1/b}) .
\]

Let \(C\) be the implicit constant in the last estimate. Then for any \(r > 2Cc(p)^{-1/b}\) it holds that

\[
P(\|X_i\| - a > r) = O(e^{-c(p)r^b/2}) .
\]

In this estimate, it is clear that if \(r = O(c(p)^{-1/b})\) then \(e^{-c(p)r^b/2} \sim 1\) while the left hand side is at most 1. Thus, the above estimate holds uniformly over \(r > 0\). Now, \((3.9)\) follows from

\[
0 \leq \mathbb{E}\|X_i\|^2 - a^2 = \mathbb{E}\|X_i\|^2 - a|^2 = O(c(p)^{-2/b}) .
\]

We obtain

\[
P(\|X_i\|^2 - 1 > r) = \begin{cases} O(e^{-c(p)r^b/C_b}), & \text{if } r \geq O(1); \\ O(e^{-c(p)r^{b/2}/C_b}), & \text{if } r = O(1). \end{cases}
\]

In particular, \((3.4)\) follows.

**Proof of \((3.3)\):** We first show that

\[
P(\|X_i + X_j\| - \mathbb{E}\|X_i + X_j\| > 2r) = O(e^{-c(p)r^b}) , \quad i \neq j ,
\]

where \(\mathbb{E}\|X_i + X_j\|\) is the high concentration case of the theorem.

\(\text{THE SPECTRUM OF RANDOM KERNEL MATRICES 11}\)
uniformly over \( r > 0 \). For any \( X \in \mathbb{R}^p \) let \( G(X) = \mathbb{E}_Y \|X + Y\| \) the expectation over \( Y \) independently sampled from the distribution of \( X_i \)'s. It is clear that

\[
\mathbb{E}_{X_i} G(X_i) = \mathbb{E}_{X_i, X_j} \|X_i + X_j\| .
\]

Using independence of \( X_i, X_j \), it follows that

\[
\begin{align*}
\text{LHS of (3.11)} \quad & \leq \mathbb{E}_{X_i} [\mathbb{E}_{X_i} [1_{\|X_i + X_j\| - G(X_i) > r}]] + \mathbb{E}_{X_i} [1_{G(X_i) - \mathbb{E}_{X_i} G(X_i) > r}] \\
& = \mathbb{E}_{X_i} O(e^{-c(p)r^b}) + O(e^{-c(p)r^\delta})
\end{align*}
\]

therefore (3.11) follows.

We now show that

\[
P(\|X_i + X_j\| - \sqrt{2} > r) = O(e^{-c(p)r^b/C_b}) ,
\]

uniformly over \( r > 0 \). Let \( \alpha = \mathbb{E}\|X_i + X_j\| \). It suffices to show that \( \alpha = \sqrt{2} + O(c(p)^{-1/b}) \). For any \( K \geq 1 \) we have

\[
\begin{align*}
\mathbb{E}\|X_i + X_j\| - \alpha^K & = \int_0^\infty Kr^{K-1}P(\|X_i + X_j\| - \alpha > r)dr \\
& = O(\mathbb{E}_X (\|X\|^K)) .
\end{align*}
\]

Letting \( K = 2 \) in the above estimate, it follows that \( \alpha = \sqrt{2} + O(c(p)^{-2/b}) \).

It follows from (3.12) that

\[
P(\|X_i + X_j\|^2 - 2 > r) = \begin{cases} 
O(e^{-c(p)r^{b/2}/C_b}), & r \geq O(1) ; \\
O(e^{-c(p)r^b/C_b}), & r = O(1) .
\end{cases}
\]

Combining this with (3.11), it follows that

\[
P(|X_i^T X_j| > r) = \begin{cases} 
O(e^{-c(p)r^{b/2}/C_b}), & r \geq O(1) ; \\
O(e^{-c(p)r^b/C_b}), & r = O(1) .
\end{cases}
\]

Consequently, for any \( K \geq 1 \) it holds that

\[
\begin{align*}
\mathbb{E}[|X_i^T X_j|^K] & = \int_0^\infty Kr^{K-1}P(|X_i^T X_j| > r)dr \\
& = \int_0^1 + \int_1^\infty = O(c(p)^{-K/b}) + O(c(p)^{-2K/b}) = O(p^{-K/2}) .
\end{align*}
\]

3.2. **Proof of Theorem 2**. For the distance model, the diagonal entries are \( f(0) \), therefore removing/adding these entries does not require any regularity of \( f \). The transference principle Lemma 1 will be used, and it remains to show that \( g(X, Y) = \|X - Y\|^2 \) satisfies the two kernel conditions of Lemma 1.

**The iid case:** We first verify (3.1). Let \( X \) and \( Y \) denote \( X_i \) and \( X_j \) for some \( i \neq j \). Then

\[
\text{Var}[\|X - Y\|^2] = -4 + \sum_{i=1}^p \mathbb{E}(x_i - y_i)^4 + \sum_{i \neq j} \mathbb{E}(x_i - y_i)^2(x_j - y_j)^2
\]

Using \( \mathbb{E}x_i^4 \) and \( \mathbb{E}y_i^4 = O(1/p^2) \), it is clear that

\[
\sum_{i=1}^p \mathbb{E}(x_i - y_i)^4 = O(1/p) .
\]
Using also independence of \( x_i, x_j, y_i, y_j \), we have
\[
\sum_{i \neq j} \mathbb{E}(x_i - y_i)^2(x_j - y_j)^2 = 4 \sum_k \mathbb{E}x_k^2 \sum_k \mathbb{E}y_k^2 + O(1/p) = 4 + O(1/p) ,
\]
and (3.1) follows.

We now verify (3.2). Fix any \( \delta > 0 \). Using the previously obtained bounds (3.4) and (3.8), it follows from the triangle inequality that
\[
\mathbb{P}(\max_{i \neq j} ||X_i - X_j||^2 - 2 \geq \delta)
\]
\[
\leq 2\mathbb{P}(\max_i ||X_i||^2 - 1 > \delta/4) + \mathbb{P}(\max_j |X_i^T X_j| > \delta/2)
\]
\[
= o(1) ,
\]
as desired.

The high concentration case: Note that (3.2) follows from (3.10) and (3.13). In fact, for any \( r > 0 \) it holds that
\[
P(||X_i - X_j||^2 - 2 > 3r) \leq 2P(||X_i||^2 - 1 > r) + P(|X_i^T X_j| > r)
\]
\[
= \begin{cases} O(e^{-c(p)^{1/2}/C}), & r > 2 ; \\ O(e^{-c(p)^{1/2}/C}), & r = O(1) . 
\end{cases}
\]
Via the same argument as before, we also obtain
\[
\mathbb{E}||X_i - X_j||^2 - 2 ||^K = O(p^{-K/2}) ,
\]
for any \( K > 2 \), and taking \( K = 2 \) gives us the first kernel condition.

3.3. Proof of Lemma 2 and Lemma 1

Proof of Lemma 2 Using (2.4), it suffices to show that \( \mathbb{E}|m_{A_1}(z + f(b)) - m_{A_2}(z)| = O(\epsilon) \) for any \( \epsilon > 0 \), which is fixed in the rest of the proof.

Let \( I_n \) denote the \( n \times n \) identity matrix. By a standard argument,
\[
|m_{A_1}(z + f(b)) - m_{A_2}(z)| \leq ||(A_1 - (z + f(b)I_n)^{-1}) - (A_2 - zI_n)^{-1}||
\]
here \( ||.|| \) denotes the spectral norm of a matrix,
\[
\leq C\|A_1 - f(b)I_n - A_2\|
\]
\[
\leq C \sup_{1 \leq i \leq n} |f(g(X_i, X_i)) - f(b)| .
\]
Since \( f \) is continuous at \( b \), there exists \( \delta = \delta(f, \epsilon) > 0 \) such that
\[
|f(x) - f(b)| < \epsilon \text{ if } |x - b| \leq \delta .
\]
Since \( m_{A_1}(z), m_{A_2}(z) = O(1) \), it follows that
\[
\mathbb{E}|m_{A_1}(z + f(b)) - m_{A_2}(z)| = \]
\[
= O(\epsilon) + O\left( P( \sup_{1 \leq i \leq n} |f(g(X_i, X_i)) - f(b)| > \delta) \right) .
\]
Therefore \( \mathbb{E}|m_{A_1}(z + f(b)) - m_{A_2}(z)| = O(\epsilon) \), thanks to (3.3). \( \square \)
Proof of Lemma 1. Without loss of generality we may assume that \( a = 0 \). Recall that \( A \) is defined using (1.2).

Let \( h(x) = f(0) + f'(0)x \), and let \( B \) be obtained from \( A \) by replacing \( f \) with \( h \) (while keeping the same kernel \( g \)). More specifically,

\[
B = f(0)M_1 + f'(0)A - f(0)I_n
\]

here \( I_n \) is the \( n \times n \) identity matrix and \( M_1 \) is the \( n \times n \) matrix whose entries are all 1’s (in particular \( M_1 = f(0)I_n + f'(0)A - f(0)I_n \)). Thus, it suffices to show that \( m_A(z) - m_B(z) \to 0 \) for any \( z \) in the upper half plane. Using (2.1), this follows from (3.14)

\[
|\mathbb{E}m_A(z) - \mathbb{E}m_B(z)| \to 0
\]

which will be shown in the rest of the proof.

Fix \( \epsilon > 0 \), it suffices to show that \( \mathbb{E}|m_A(z) - m_B(z)| = O(\epsilon) \) for \( n, p \) sufficiently large. Let \( \lambda_1(A) \leq \cdots \leq \lambda_n(A) \) be the eigenvalues of \( A \) and \( \lambda_1(B) \leq \cdots \leq \lambda_n(B) \) be the eigenvalues of \( B \). Then for any fixed \( z \notin \mathbb{R} \) it holds that

\[
\frac{|m_A(z) - m_B(z)|^2}{\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lambda_i(A) - z} - \frac{1}{\lambda_i(B) - z}} \leq \text{(Cauchy-Schwarz)}
\]

\[
\leq C \frac{1}{n} \sum_{i=1}^{n} |\lambda_i(A) - \lambda_i(B)|^2 \quad \text{(using Im(z) > 0)}
\]

where \( C > 0 \) is a constant which may depend on \( z \). It then follows from the Hoffman–Wielandt inequality (see e.g. [33]) that

\[
|m_A(z) - m_B(z)|^2 \leq C \frac{1}{n} \sum_{i \neq j} |A_{ij} - B_{ij}|^2.
\]

By definition, there is \( \delta > 0 \) depending on \( f \) only such that

\[
|f(x) - h(x)| \leq \epsilon |x| \quad \text{for } |x| < \delta.
\]

Let \( F \) be the event that there is a pair \( i \neq j \) such that \( |g(X_i, X_j)| \geq \delta \). It follows from the second assumption (3.22) on \( g \) that \( \mathbb{P}(F) \leq \epsilon^2 \) for \( n \) (and \( p \)) sufficiently large. We now estimate

\[
\mathbb{E}[|m_A(z) - m_B(z)|^2] \leq \mathbb{E}[1_F |m_A(z) - m_B(z)|^2] + \mathbb{E}[1_{\mathbb{C}^2} |m_A(z) - m_B(z)|^2].
\]

Since \( m_A(z) = O(1) \) and \( m_B(z) = O(1) \), it follows that

\[
\mathbb{E}[1_F |m_A(z) - m_B(z)|^2] = O(\mathbb{P}(F)) = O(\epsilon^2)
\]

for large \( n \) large \( p \). On the other hand, from (3.16) and (3.17), it follows that

\[
\mathbb{E}[1_{\mathbb{C}^2} |m_A(z) - m_B(z)|^2] = O(\frac{1}{n} \sum_{i \neq j} \mathbb{E}(\epsilon^2 |g(X_i, X_j)|^2)).
\]

With the first assumption on \( g \), it follows that

\[
\mathbb{E}[1_{\mathbb{C}^2} |m_A(z) - m_B(z)|^2] = O(\epsilon^2).
\]

Consequently, in the large \( n \) large \( p \) limit it holds that

\[
\mathbb{E}|m_A(z) - m_B(z)| \leq (\mathbb{E}|m_A(z) - m_B(z)|^2)^{1/2} = O(\epsilon).
\]

This completes the proof of (3.14).
3.4. Concentration of the Stieltjes transform. In this section, we include a proof of (2.1).

Lemma 3. [13] Let $M$ be an $p \times n$ matrix with independent entries. Let $A$ be defined by $A_{ij} = F(M_i, M_j, p)$ any real-valued function that is symmetric in the first two variables, here $M_i$ denotes the $i$th column of $M$. Let $\hat{A}$ have the same non-diagonal entries as $A$, and zero diagonal entries. Fix $z$ with $\text{Im}(z) > 0$. Then

\begin{align}
(3.18) \quad & \lim_{n \to \infty} |m_A(z) - \mathbb{E}m_A(z)| = 0 \quad \text{a.s.} \\
(3.19) \quad & \lim_{n \to \infty} |m_{\hat{A}}(z) - \mathbb{E}m_{\hat{A}}(z)| = 0 \quad \text{a.s.}
\end{align}

Proof. The proof largely follows the argument in [13], which is a variant of standard arguments (see e.g. [5]). Only (3.18) will be proved below, the proof for (3.19) is entirely similar. Using Borel-Cantelli’s lemma, it suffices to show

\begin{equation}
(3.20) \quad \mathbb{E}|m_A(z) - \mathbb{E}m_A(z)|^4 = O(n^{-2}) .
\end{equation}

Recall the following standard estimate (i.e. Khintchine’s inequality) for the martingale square function:

\begin{equation}
(3.21) \quad \mathbb{E} \sum_{j=1}^{n} \Delta_j|^p \leq C_p \mathbb{E}\left( \sum_{j=1}^{n} |\Delta_j|^2 \right)^{p/2} , \quad 1 < p < \infty ,
\end{equation}

where $\Delta_j = S_{j+1} - S_j$ the martingale difference sequence. Below, a martingale will be constructed such that $S_0 = \mathbb{E}m_A(z)$ and $S_n = m_A(z)$, and then show that the corresponding right hand side of (3.21) is bounded above by a suitable power of $n$.

For any $0 \leq k \leq n$ let $\mathcal{A}_k$ be the sigma algebra generated by the last $k$ columns of $M$. Then define $S_k := \mathbb{E}_k \text{Tr}(A - z)^{-1}$ which is a martingale with respect to the filtration $\{\mathcal{A}_0 \subset \cdots \subset \mathcal{A}_n\}$ . Since $m_A(z)$ is measurable with respect to $\mathcal{A}_n$, the construction gives $S_n = nm_A(z)$ while clearly $S_0 = n\mathbb{E}m_A(z)$. It follows from (3.21), with $p = 4$, that

\[ \mathbb{E}|m_A(z) - \mathbb{E}m_A(z)|^4 \leq \frac{C}{n^2} \mathbb{E}\left( \max_{1 \leq j \leq n} |\Delta_j|^4 \right) . \]

It therefore suffices to show that, uniformly over $j$

\begin{equation}
(3.22) \quad \mathbb{E}_j |\text{Tr}(A - z)^{-1} - \mathbb{E}_{j-1} |\text{Tr}(A - z)^{-1}| = O(1)
\end{equation}

where the implicit constant is allowed to depend on $z$. Fix $1 \leq j \leq n$ and let $B$ denote the $(n - 1) \times (n - 1)$ submatrix of $A$, obtained by deleting the $j$th row and the $j$th column of $A$. By definition of $A$, the entries of $B$ are independent of $M_j$, therefore $\mathbb{E}_j B = \mathbb{E}_{j-1} B$ and consequently it suffices for (3.22) to show that

\[ |\text{Tr}(A - z)^{-1} - \text{Tr}(B - z)^{-1}| = O(1) . \]

Since $f$ is symmetric and real-valued, the eigenvalues of $A$ and $B$ are real valued. Furthermore, they interlace by the Cauchy interlacing theorem (see e.g. [33]). The desired estimate now follows immediately:

\[ |\text{Tr}(A - z)^{-1} - \text{Tr}(B - z)^{-1}| \leq \int \frac{d}{dt} \left( \frac{1}{t-z} \right) dt = O\left( \frac{1}{\text{Im}(z)} \right) . \]

\[ \square \]
4. The $p$-dependent setting

4.1. Some estimates for orthogonal polynomials. In this section, some basic estimates involving orthogonal polynomials are proved, and these estimates will be used in the proof of Theorem 3.

Let $h_k(x)$ denote the $k$th Hermite polynomial, i.e. the orthogonal polynomial with respect to the Gaussian measure $\mu(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

Let $p_k,p$ denote the $k$th orthogonal polynomial with respect to $\mu_p$, the probability distribution of $\xi_p = \sqrt{p}Z_1^T Z_2$ where $Z_1, Z_2$ are iid random $1 \times p$ vectors whose coordinates are independently sampled from a random variable $Z$ satisfying

$$E[Z] = 0, \quad \text{Var}[Z] = 1/p,$$

and for some $K > 0$ sufficiently large

$$E[|Z|^K] = O(p^{-K/2}).$$

**Lemma 4.** Let $k \geq 0$. If (4.1) holds for $K$ sufficiently large then for any $\delta > 0$ it holds that

$$|p_{k,p}(x) - h_k(x)| \leq C\delta(1 + |x|^k), \quad x \in \mathbb{R},$$

where the implicit constant $C$ may depend on $k$ (but not on $p$).

**Proof.** For $k \geq 0$ let $m_k = E\xi_p^k$. Then for some normalization constant $c_k$ it holds that (see e.g. [32]):

$$p_{k,p}(x) = c_k \det \begin{pmatrix} m_0 & m_1 & \ldots & m_k \\ \vdots & \vdots & \ddots & \vdots \\ m_{k-1} & m_k & \ldots & m_{2k-1} \\ 1 & x & \ldots & x^k \end{pmatrix}.$$ 

The leading coefficient of $p_{k,p}$ is $c_k \det M_{k-1}$ where

$$M_j = \begin{pmatrix} m_0 & \ldots & m_j \\ \vdots & \ddots & \vdots \\ m_j & \ldots & m_{2j} \end{pmatrix}.$$ 

Since $\|p_{k,p}\|_{L^2(\mu_p)} = 1$, it follows that

$$c_k^2 = \frac{1}{(\det M_{k-1})(\det M_k)}.$$ 

and the sign of $c_k$ is completely determined from the sign of $\det M_{k-1}$. Therefore in order to show (4.2) it suffices to show that if $\delta > 0$ then

$$E\xi_p^k = EN^k + O_{k}(\delta),$$

where $N$ is the normal Gaussian $\mathcal{N}(0,1)$. But this is a classical theorem of von Bahr [37].

**Lemma 5.** Let $a_{j,p}$ be the coefficients in the orthogonal expansion of a normalized kernel function $k$ satisfying $E[k(\xi_p,p)^2] = O(1)$ uniformly over $p$. Then

$$|a_{j,p}| = O(1).$$
Proof. Clearly for any $j$
\[ |a_{j,p}| = |\int k(x,p)p_j(x)d\mu_p(x)| \]
\[ \leq \|k(\cdot,p)\|_{L^2(\mu_p)}\|p_j,p\|_{L^2(\mu_p)} = O(1) \] .

Note that if the kernel $k$ satisfies (1.8) then Lemma 5 applies.

4.2. Proof of Theorem 3
Let $M_A$ be the $p \times n$ matrix whose columns are $X_1, \ldots, X_n$. Let $M_G$ denote the $p \times n$ matrix whose columns are iid random Gaussian vectors $G_1, \ldots, G_n$, which are normalized so that the entries of $M_G$ have mean 0 and variance $1/p$. A Gaussian analogue $G$ of $A$ will be constructed as follows. The construction will ensure, thanks to [13], that the spectral density of $G$ converges to the desired limiting spectral density in Theorem 3.

Let $\xi_{G,p} = \sqrt{p}G_1^T G_2$ and let $\mu_{G,p}$ denote its probability distribution. Let
\[ K(x,p) = \sum_{i=0}^{\infty} a_{i,p} P_{i,p}(x) \]
where $P_{i,p}$ is the $i^{th}$ orthogonal polynomial with respect to $\mu_{G,p}$. Note that the above infinite sum converges uniformly in $L^2(\mu_{G,p})$ by given assumptions on $a_{i,p}$.

Let $F(x,p) = (\sqrt{p})^{-1}K(x,\sqrt{p},p)$ and let $G$ be the random matrix generated from $M_G$ using $F$. It was shown in [13] that the Stieltjes transform $m_G(z)$ of $G$ converges pointwise to the solution of (1.10). Therefore, using (2.1), it suffices to show that in the large $n$ large $p$ limit it holds that
\[ E[m_A(z) - m_G(z)] = O(\epsilon) \]
(4.4) where $\epsilon > 0$ is fixed in the rest of this section.

It follows from the given assumption (1.7) that there exists $L = L(\epsilon)$ such that uniform over large $p$ it holds that $\sum_{i > L} |a_{i,p}|^2 = O(\epsilon^2)$. Let
\[ k_L(x,p) := \sum_{h=0}^{L} a_{h,p} p_{h,p}(x) \]
and
\[ K_L(x,p) := \sum_{h=0}^{L} a_{h,p} P_{h,p}(x) \] .
We obtain
\[ E[|k(\xi_p,p) - k_L(\xi_p,p)|^2] = O(\epsilon^2) \]
(4.5)
\[ E[|K(\xi_G,p) - K_L(\xi_G,p)|^2] = O(\epsilon^2) \] .
(4.6)
Let $f_L$ and $F_L$ correspond to $k_L$ and $K_L$, and let $A_L$ and $G_L$ be generated from them respectively.

It follows from (4.5) and (4.6) and Lemma [8] below that
\[ E[m_A(z) - m_{A_L}(z)] = O(\epsilon) \]

\[ E[m_G(z) - m_{G_L}(z)] = O(\epsilon) \] .
The following Lemma is in [13], to keep the paper self-contained, a short proof of Lemma [8] is included.
Lemma 6. Let $Y$ a random vector of length $p$ whose coordinates are iid with mean 0 and variance $1/p$. Let $Y'$ be an iid copy of $Y$. Assume that for $p$ large

$$\mathbb{E}|f_1(Y^TY', p) - f_2(Y^TY', p)|^2 \leq \epsilon^2/p.$$ 

Let $A_1$ and $A_2$ be generated from $f_1$ and $f_2$ using $n$ iid copies of $Y$. Then in the large $p$ large $n$ limit it holds that

$$\mathbb{E}|m_{A_1}(z) - m_{A_2}(z)| = O(\epsilon).$$

Proof. We largely follow [13]. Following the proof of (3.16), it is clear that

$$\mathbb{E}[m_{A_1}(z) - m_{A_2}(z)]^2 \leq Cn^{-1}\sum_{i \neq j} [A_{1,ij} - A_{2,ij}]^2$$

$$= Cn^{-1}(n-1)[f_1(Y^TY', p) - f_2(Y^TY', p)]^2$$

$$= O(\epsilon^2),$$

which implies the desired estimate. \hfill $\square$

Thus, it suffices for [13] to show that $\mathbb{E}[m_{A_L}(z) - m_{G_L}(z)] = O(\epsilon)$. This will be proved in two steps. First, it will be shown in Section 4.3 that

$$\mathbb{E}[m_{A_L}(z) - m_{\tilde{G}_L}(z)] = O(\epsilon),$$

where $\tilde{G}_L$ is generated from $M_L$ using $k_L$ as opposed to $K_L$, which was used to generate $G_L$. Then in Section 4.4 it will be shown that

$$\mathbb{E}[m_{G_L}(z) - m_{\tilde{G}_L}(z)] \to 0.$$

4.3. Proof of (4.7): Conversion from $A_L$ to $\tilde{G}_L$. The proof of (4.7) will follow the strategy in [14]: the idea is to convert $A_L$ to $\tilde{G}_L$ in $np$ steps, in each step one entry in $M_A$ is replaced by the corresponding entry in $M_G$. It suffices to show that in every step it holds that $\mathbb{E}[\Delta m(z)] = O(n^{-2}\epsilon)$, here $\Delta m$ is the difference between the Stieltjes transforms of the underlying matrices. This is the content of Lemma 7.

The generic setting for each step can be formulated as follows. Let $M[1]$ and $M[2]$ denote two $p \times n$ random matrices that share the same entries except for the $(i, j)$ position. Assume that these entries are independent, and their distribution have mean 0, variance $1/p$, and higher moments bounded (with uniform constants) by properly scaled powers of $p$. Let $A[1], A[2]$ be generated from $M[1], M[2]$ using the kernel function $f_L$.

Lemma 7. In the large $n$ large $p$ limit

$$\mathbb{E}[m_{A[1]}(z) - m_{A[2]}(z)] = O(n^{-2}\epsilon).$$

For simplicity of notation, in the rest of this section $M[0]$ denotes the matrix that shares the same entries with $M[1], M[2]$ except for the $(i, j)$ position, where $M[0]_{ij} := 0$. Denote by $A[0]$ the kernel matrix generated from $M[0]$ using $f_L$.

Proof. Let $E$ be the event that $\|A[0] - A[m]\| \leq Im(z)/2$ for both $m = 1, 2$. In Lemma 8 it will be shown that

$$\mathbb{E}[\|A[0] - A[m]\|^q] = O(n^{-q/2}), \quad m = 1, 2,$$
here \( \| \cdot \| \) denotes the matrix norm. Since \( q \) could be taken large \((q > 4 \text{ suffices})\), it follows that \( E^c \) has probability \( o(n^{-2}) \). Clearly \( m_{A[0]}(z) = O(1) \), thus it remains to show
\begin{equation}
\mathbb{E}\left( 1_E |m_{A[1]}(z) - m_{A[2]}(z)| \right) = O(n^{-2}e)
\end{equation}
in the large \( n \) large \( p \) limit.

Let \( A \) denote \( A[1] \) or \( A[2] \) in the rest of the proof, and let \( M \in \{ M[1], M[2] \} \) be the corresponding sample matrix. On \( E \), expand
\begin{equation}
(A - zI)^{-1} = \sum_{k \geq 0} R_k , \quad \text{where}
\end{equation}
\[ R_k := [(A[0] - zI)^{-1}(A[0] - A)]^k(A[0] - zI)^{-1} . \]
In particular, \( R_0 = (A[0] - zI)^{-1} \). Note that any eigenvalue of \( A[0] - Iz \) is of the form \( \lambda - z \) for some eigenvalue \( \lambda \) of \( A[0] \). Since any such \( \lambda \) is real, it follows that
\begin{equation}
\|R_0\| \leq \frac{1}{Im(z)} ,
\end{equation}
Thus on \( E \) the expansion \( \text{4.11} \) is absolutely convergence with respect to \( \| \cdot \| \).

This expansion will be used to compute the leading ‘asymptotics’ of \( \mathbb{E}[1_E m_A(z)] \) as \( n \to \infty \), and show that modulo \( o(n^{-2}) \) one obtains the same leading asymptotics for \( A = A[1] \) and \( A = A[2] \), and this clearly implies \( \text{4.10} \).

In \( \text{4.11} \), it is clear that the contributions to \( \mathbb{E}[1_E m_A(z)] \) of \( R_0 \) is the same for \( A = A[1] \) and \( A = A[2] \), so below only \( R_k \) with \( k \geq 1 \) is considered.

### Decay estimates for contribution of higher order terms in \( \text{4.11} \):
First, it will be shown that for some absolute constant \( C \) the following holds for \( k \geq 1 \):
\begin{equation}
|\text{Tr}(R_k)| \leq CIm(z)^{-(k+1)}\|A[0] - A\|^k
\end{equation}
To show \( \text{4.13} \), the key observation is that \( A[0] - A \) has rank at most 2. Indeed, this follows from the fact that at most one column and at most one row in \( A[0] - A \) could be nonzero. Thus, \( [R_0(A[0] - A)]^kR_0 \) is of rank at most 2 and therefore
\begin{equation}
|\text{Tr}(R_k)| \leq 2\|R_0(A[0] - A)]^kR_0\| \\
\leq 2\|R_0\|^{k+1}\|A[0] - A\|^k ,
\end{equation}
and \( \text{4.13} \) follows immediately from \( \text{4.14} \) and \( \text{4.12} \).

As a consequence of \( \text{4.13} \), it follows that
\begin{equation}
n^{-1}\mathbb{E}\left[ 1_E \sum_{k \geq 3} \text{Tr}R_k \right] \leq Cn^{-1}\mathbb{E}[\|A[0] - A\|^3] = O(n^{-5/2}) .
\end{equation}

### Asymptotics matching for the contribution of \( R_1 \):
Rewrite
\begin{equation}
n^{-1}\mathbb{E}[1_E\text{Tr}(R_1)] = n^{-1}\mathbb{E}[\text{Tr}(R_1)] - n^{-1}\mathbb{E}[1_E\text{Tr}(R_1)] .
\end{equation}
We first show that the second term in \( \text{4.15} \) is \( o(n^{-2}) \). Indeed, it follows from Cauchy-Schwarz that
\begin{equation}
n^{-1}|\mathbb{E}[1_E\text{Tr}(R_1)]|
\end{equation}
where \(a, b, c\), which implies the desired estimate.

It will be shown in Lemma 9 that there exists a decomposition (4.16)

\[ A[0] - A = aM_{ij} + bM_{ij}^2 + c, \]

where \(a, b, c\) are \(n \times n\) matrices such that the following holds:

- The entries outside the \(j^{th}\) rows and the \(i^{th}\) columns of \(a, b, c\) are zeros.
- The nonzero entries of \(a\) and \(b\) depend only on \(f\) and \(M[0]\).
- The \(2^{nd}\) moment of any entries of \(a\) and \(b\) are of size \(O(n^{-1})\), and the \(2^{nd}\) moment of any entries of \(c\) are of size \(O(n^{-4})\).

It follows that \(E[\|c\|^2] = O(n^{-3})\) and \(c\) has rank 2, hence

\[ \text{Tr}[E_0 R_0 E^{(ij)} c R_0] = \text{Tr}[R_0 c R_0] = O(\|c\|) = O(n^{-3/2}) \]

It follows that

\[ n^{-1}E[\text{Tr}(R_1)] = n^{-1}E_0 \left( \text{Tr} \left[ R_0 a R_0 \right] \right) E^{(ij)} [M_{ij}] + n^{-1}E_0 \left( \text{Tr} \left[ R_0 b R_0 \right] \right) E^{(ij)} [M_{ij}^2] + O(n^{-5/2}) \]

We remark that the first two terms are the same for \(A = A[1]\) and \(A = A[2]\).

**Asymptotics matching for the contribution of \(R_2\):**

As before, rewrite

\[ n^{-1}E[1_E \text{Tr}(R_2)] = n^{-1}E[\text{Tr}(R_2)] - n^{-1}E[1_E \text{Tr}(R_2)] \]

and using (4.14) it is not hard to see that the second term is \(o(n^{-2})\). Therefore it remains to compute the asymptotics for the first term. Again, the decomposition (4.16) will be used to expand

(4.17) \[ n^{-1}E[\text{Tr}(R_2)] = n^{-1}E^{(ij)} [M_{ij}^2] E_0 \left( \text{Tr}[(R_0 a)^2 R_0] \right) \]

Since the second moment of any entries of \(a\) and \(b\) are \(O(n^{-1})\) and since \(a\) and \(b\) have \(O(n)\) non-zero entries, it is not hard to see that \(E[\|a\|^2] = O(1)\) and \(E[\|b\|^2] = O(1)\), while \(E[\|c\|^2] = O(n^{-3})\) as remarked above. Therefore using the small rank properties of \(a, b, c\), the other terms in the expansion (4.17) can be bounded by expected values of products of spectral norms, and eventually obtain an estimate of \(O(n^{-5/2})\). Clearly, the first term in the expansion (4.17) is the same for \(A = A[1]\) and \(A = A[2]\).

Finally, to complete the proof of Lemma 7 it remains to show (4.9) (see Lemma 8) and (4.16) (see Lemma 9). \(\square\)
To keep the following Lemmas self-contained, the symbol \( H[k] \) will be used instead of \( A[k] \). In the applications of these Lemma to obtain (4.9) and (4.16), these are the same.

**Lemma 8.** For each \( 0 \leq k \leq 2 \) let \( H[k] \) be generated from the column vectors of \( M[k] \) using a kernel function \( h(x,p) \) using the non-diagonal model (1.2). Assume that

\[
|\partial_x h(x,p)| \leq C_N (1 + |\sqrt{px}|^N),
\]

for some \( N > 0 \) uniform over \( p \) and \( x \). Then for \( m = 1, 2 \) and \( q > 0 \) it holds that

\[
\mathbb{E} \left( \|H[0] - H[m]\|^q \right) = O_N (p^{-q/2}).
\]

Remark: To obtain (4.9), Lemma 8 will be applied for \( h(x,p) = f_L(x,p) \), which satisfies (4.18) thanks to Lemma 4 and Lemma 5. Note that in that case it holds that \( H[k] = A[k] \).

**Proof.** For simplicity, let \( H = H[m] \) and \( M = M[m] \) and \( U = M[0]^T M[0] \). Since \( H[0] - H \) has at most \( O(n) \) nonzero entries and since

\[
\|B\| \leq \left( \sum_{i,j} |B_{ij}|^2 \right)^{1/2}
\]

for any square matrix \( B \), it suffices to show that the \( q \) moment of each nonzero entry of \( H[0] - H \) is bounded above by \( O(n^{-q}) \).

Consider without loss of generality an (non-diagonal) entry on the \( j^{th} \) row of \( H[0] - H \). This entry has the form

\[
\alpha_{jk} = h(U_{jk},p) - h(U_{jk} + M_{ij}M_{ik},p)
\]

for some \( 1 \leq k \leq n \). It follows from (4.18) that

\[
|\alpha_{jk}| = O_N \left( |M_{ij}M_{ik}|(1 + \sqrt{p} |M_{ij}M_{ik}| + \sqrt{p} |U_{jk}|)^N \right)
\]

Using (1.6), it is not hard to see that \( \mathbb{E} [\|\sqrt{p} U_{jk}\|^K] = O(1) \) (for details see for instance the proof of (4.8)). Also, it is clear that \( \mathbb{E} [pM_{ij}M_{ik}]^K = O(1) \). Therefore, using Cauchy-Schwarz it follows that

\[
\mathbb{E} \left[ |\alpha_{jk}|^q \right] \leq C_N \left( \mathbb{E} |M_{ij}M_{ik}|^{2q} \right)^{1/2} = O(p^{-q}).
\]

□

**Lemma 9.** For each \( 0 \leq k \leq 2 \) let \( H[k] \) be generated from the column vectors of \( M[k] \) using a kernel function \( h(x,p) \) using the non-diagonal model (1.2). Assume that for each \( m = 1, 2, 3 \) there is some \( N > 0 \) such that

\[
|\partial_x^m h(x,p)| \leq C_N p^{(m-1)/2} (1 + |\sqrt{px}|^N),
\]

where \( C_N \) is uniform over \( p \) and \( x \). Then for \( m = 1, 2 \) it holds that

\[
H[m] - H[0] = aM[m]_{ij} + b \left[ M[m]_{ij} \right]^2 + c,
\]

where \( a, b, c \) are \( n \times n \) matrices such that

(i) the entries outside the \( j^{th} \) rows and columns of \( a, b, c \) are zeros, and

(ii) the nonzero entries of \( a \) and \( b \) are independent of \( M[m]_{ij} \), and
(iii) the 2nd moment of any entries of a and b are of size $O(n^{-1})$, and the 2nd moment of any entries of c are of size $O(n^{-4})$.

Remark: The condition (4.19) is satisfied when $h = f_L$ thanks to Lemma 4 and Lemma 5. Thus, (4.16) follows by applying Lemma 9 to $h = f_L$, in that case it holds that $H[k] = A[k]$.

Proof. Let $H = H[m]$ and $M = M[m]$ for $m \in \{1, 2\}$. Let $U = M[0]^T M[0]$. It is clear that all nonzero entries of $H - H[0]$ must be in the $j^{th}$ column or the $j^{th}$ row and off the diagonal, thus it remains to decompose these non-zero entries.

Consider without loss of generality an entry in the $j^{th}$ row of $H - H[0]$, which has the form

$$
\alpha_{jk} = h(U_{jk} + M_{ij} M[0]_{ik}, p) - h(U_{jk}, p)
$$

for some $k \neq j$. We now decompose

$$
\alpha_{jk} = a_{jk} M_{ij} + b_{jk} M_{ij}^2 + c_{jk}
$$

$$
a_{jk} = M[0]_{ik} h_x(U_{jk}, p), \quad b_{jk} = \frac{1}{2} \left[ M[0]_{ik} \right]^2 h_{xx}(U_{jk}, p).
$$

Similar decomposition for the $j^{th}$ column of $A[0] - A$ also holds. It is clear that the two conditions (i) and (ii) are satisfied and it remains to show that

$$
\mathbb{E}|c_{jk}|^2 = O(n^{-4}).
$$

By given assumption, for some $\theta$ in between $U_{jk}$ and $U_{jk} + M_{ij} M[0]_{ik}$ it holds that

$$
c_{jk} = \frac{1}{6} \left[ M_{ij} M[0]_{ik} \right]^3 h_{xxx}(\theta, p).
$$

In particular, it follows from (4.19) that, for some $N > 0$ and $C$ depends on $N$,

$$
|c_{jk}| \leq C p |M_{ij} M[0]_{ik}|^3 \left[ 1 + (\sqrt{p}\theta)^N \right],
$$

$$
\leq C p |M_{ij} M[0]_{ik}|^3 \left[ 1 + p^{N/2} (|U_{jk}| + |M_{ij} M[0]_{ik}|)^N \right].
$$

Using Hölder’s inequality and the previously obtained bounds on moments of entries of $M, M[0], U$, it follows that

$$
\mathbb{E}|c_{jk}|^2 = O(p^{-4}).
$$

\[\square\]

4.4. Proof of (4.3): Conversion from $\bar{G}_L$ to $G_L$. Recall that $\xi_{G,p} = \sqrt{p} Y Y'$ where $Y$ and $Y'$ denote iid $1 \times p$ vectors with independent Gaussian $N(0, 1/p)$ coordinates. Applying Lemma 9 it remains to show that: for any $\delta > 0$, the following holds uniformly over $p$ large

$$
\mathbb{E} \left[ \left| k_L(\xi_{G,p}, p) - K_L(\xi_{G,p}, p) \right|^2 \right] = O(\delta),
$$

here the implicit constant depends on $L$.

It follows from the triangle inequality and Lemma 11 that, for $p$ large,

$$
|p_{k,p}(x) - P_{k,p}(x)| \leq C_k \delta (1 + |x|^k).
$$

On the other hand it is clear that $E|\xi_p|^k = O_k(1)$. It follows that, for $p$ large,

$$
\|p_{k,p} - P_{k,p}\|_{L2(\mu_{G,p}, \mathbb{R})} \leq C_k \delta
$$
for each \( k = 0, L \). Now (4.20) follows since
\[
K_L(x, p) - k_L(x, p) = \sum_{k=1}^{L} a_{k,p} \left[ p_{k,p}(x) - p_{k,p}(x) \right],
\]
and for \( p \) large \( a_{k,p} = O(1) \) by Lemma 6.

5. CONCLUDING REMARKS

In Theorem 3 it was assumed that the entries of \( X_i \)'s are i.i.d., while in Theorem 1 and Theorem 2 it is possible to have random vectors with dependent entries, as long as a high concentration condition is satisfied. Cheng and Singer 13 on the other hand have outlined a proof of an analogue of Theorem 3 in the setting when \( X_i \)'s are independently identically sampled from the unit sphere. This suggests that the i.i.d. assumption on entries of \( X_i \)'s may be weakened. Our proof of Theorem 3 in this paper however relies on the independence of entries of \( X_i \), more specifically in the implementation of the Lindeberg swapping argument in Section 4.3. It would be interesting to see if this swapping argument could be improved to extend Theorem 3 to settings when \( X_i \)'s have dependent entries. In this direction, see for instance Chatterjee 11 where some generalization of the Lindeberg principle was considered.

In a different direction, one may ask questions about local statistics of the eigenvalues of random kernel matrices. The local statistics of Wigner and covariance matrices have been studied extensively in the literature, see e.g. 34 or 35 for a comprehensive survey. However, we are not aware of any related work in the setting of random kernel matrices, even when the envelop function is independent of \( p \). It seems that a naive adaptation of the approximation argument, carried out in this paper and El Karoui’s work 15 16 17, does not lead to sufficiently interesting information about local statistics of the eigenvalues, unless very special assumptions are made on the kernel. El Karoui 15 16 17 on the other hand has been able to obtain some results about behavior of the largest eigenvalues via the approximation approach.

REFERENCES

[1] Adamczak, Radoslaw (2011). On the Marchenko–Pastur and circular laws for some classes of random matrices with dependent entries. Electron. J. Probab. 16 no. 37 1068–1095. MR 2820070
[2] Andrews, George E., Askey, Richard, and Roy, Ranjan (2001). Special Functions (Encyclopedia of Mathematics and its Applications). Cambridge University Press, Cambridge.
[3] Aubrun, G. (2006). Random points in the unit ball of \( l_n^p \). Positivity no. 10(4) 755–759.
[4] Bai, Z. D. (1999). Methodologies in spectral analysis of large-dimensional random matrices, a review. Statist. Sinica 9 611–677 (1999).
[5] Bai, Zhidong and Silverstein, Jack W. (2010). Spectral analysis of large dimensional random matrices, 2nd ed. Springer Series in Statistics. Springer, New York. MR 2567175 (2011d:60014)
[6] Bogomolny, E., Bohigas, O., and Schmit, C. (2003). Spectral properties of distance matrices. J. Phys. A (2003) 36 no. 12 3595–3616. MR 1986436 (2004e:15027)
[7] Bordenave, B., Caputo, P., and Chafa, D. (2011). Circular law theorem for random Markov matrices. Probability Theory and Related Fields.
[8] Bordenave, Charles (2008). Eigenvalues of euclidean random matrices. Random Structures Algorithms 33 33 no. 4, 515532. MR 2462254 (2009j:15082)
[9] Bordenave, Charles (2012). On Euclidean random matrices in high dimension. Preprint, arXiv:1209.5888 [math.PR].
10. Chaia, Djalil (2010). The Dirichlet Markov ensemble. J. Multivariate Anal. 101 no. 3 555–567. MR 2575404 (2010m:60019)
11. Chatterjee, Sourav (2006). A generalization of the Lindeberg principle. Ann. Probab. 34 no. 6 2061–2076. MR 2294926 (2008c:60028)
12. Chatterjee, Sourav, Diaconis, Persi and Sly, Allan (2010). Properties of Uniform Doubly Stochastic Matrices. Preprint, arXiv:1010.6136v1 [math.PR].
13. Cheng, Xiuyuan and Singer, Amit (2012). The spectrum of random inner-product kernel matrices. Preprint, arXiv:1202.3155 [math.PR].
14. Durrett, Rick (2010). Probability: theory and examples, 4th ed. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge. MR 2722836 (2011e:60001)
15. El Karoui, Noureddine (2007). Concentration of measure and spectra of random matrices: with applications to correlation matrices, elliptical distributions and beyond. The Annals of Applied Probability, to appear.
16. El Karoui, Noureddine (2007). Operator norm consistent estimation of large dimensional sparse covariance matrices. The Annals of Statistics, to appear.
17. El Karoui, Noureddine (2010). The spectrum of kernel random matrices. Ann. Statist. 38 no. 1 1–50. MR 2589315 (2011a:62187)
18. El Karoui, Noureddine (2012). On information plus noise kernel random matrices. The Annals of Statistics, to appear.
19. Götze, F. and Tikhomirov, A. (2004). Limit theorems for spectra of positive random matrices under dependence. Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) 311 Veroyatn. i Stat. 5 121–124, 299. (English, with English and Russian summaries); English transl. in J. Math. Sci. (N. Y.) 133 (2006) no. 3 1257–1276. MR 2092202 (2005m:15067)
20. Götze, F. and Tikhomirov, A. (2005). Limit theorems for spectra of random matrices with martingale structure. Stein’s method and applications. Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap. 5 181–193. Singapore Univ. Press, Singapore. MR 2205336 (2007a:60037)
21. Guédon, Olivier and Milman, Emanuel (2011). Interpolating thin-shell and sharp large-deviation estimates for isotropic log-concave measures. Geom. Funct. Anal. 21 no. 5 1043–1068. MR 2846382
22. Jiang, Tiefeng (2012). Distributions of Eigenvalues of Large Euclidean Matrices Generated from Three Manifolds. Preprint.
23. Kolchin, V. and Giné, E (2000). Random matrix approximation of spectra of integral operators. Bernoulli 6 113–167.
24. Ledoux, Michel (2001). The concentration of measure phenomenon. Mathematical surveys and monographs. American Math. Society, Providence, RI.
25. Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues in certain sets of random matrices. Mat. Sb. (N. Y.) 72 (114) 507–536. MR0206649 (34 #8458).
26. Mélard, M., Parisi, G., and Zee, A. (1999). Spectra of Euclidean random matrices. Nuclear Phys. B 559 (1999) no. 3 689–701. MR 1724455 (2000h:82004)
27. O’Rourke, Sean (2012). A note on the Marchenko-Pastur law for a class of random matrices with dependent entries. Preprint, arXiv:1201.3854.
28. Pajor, A. and Pastur, L. (2009). On the limiting empirical measure of eigenvalues of the sum of rank one matrices with log-concave distribution. Studia Math. 195 no. 1 11–29. MR 2539559 (2010h:15022
29. Paouris, G. (2006). Concentration of mass on convex bodies. Geom. Funct. Anal. 16 no. 5 1021–1049. MR 2276533 (2007k:52009)
30. Rasmussen, Carl Edward and Williams, Christopher K. I. (2006). Gaussian processes for machine learning. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA. MR 2514435 (2010i:68131)
31. Silverstein, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. J. Multivariate Anal. 55 no. 2 331–339. MR 1370408 (96m:60078)
32. Szegő, G. (1975). Orthogonal Polynomials, 4th ed. Amer. Math. Soc., Providence, RI.
33. Tao, Terence (2012). Topics in random matrix theory. Graduate Studies in Mathematics 132. American Mathematical Society, Providence, RI. MR 2906465
34. Tao, Terence and Vu, Van (2012). Universality of the local eigenvalues statistics. Acta Math. (to appear).
[35] Tao, Terence and Vu, Van (2012). Random matrices: The Universality phenomenon for Wigner ensembles. Preprint, arXiv:1202.0068.

[36] Vershik, A. M. (2004). Random metric spaces and universality. (Russian, with Russian summary). Uspekhi Mat. Nauk 59 (2004) no. 2(356) 65–104. translation in Russian Math. Surveys 59 (2004) no. 2 259–295. MR 2086637 (2005g:60019)

[37] von Bahr, Bengt (1965). On the Convergence of Moments in the Central Limit Theorem. Ann. Math. Statist. 36 no. 3 808–818.

[38] Wachter, Kenneth W. (1978). The strong limits of random matrix spectra for sample matrices of independent elements. Ann. Probability 6 no. 1 1–18. MR 0467894 (57 #7744)

[39] Williams, C. and Seeger, M. (2000). The effect of the input density distribution on kernel-based classifiers. International Conference on Machine Learning 17, 1159–1166.

[40] Yin, Y. Q. and Krishnaiah, P. R. (1985). Limit theorem for the eigenvalues of the sample covariance matrix when the underlying distribution is isotropic. Teor. Veroyatnost. i Primenen. 30 no. 4 810–816. MR 816299 (87m:60060)

E-mail address: yen.do@yale.edu

E-mail address: van.vu@yale.edu

Department of Mathematics, Yale University, New Haven, CT 06520-8283