THE SPECTRAL NORM OF RANDOM INNER-PRODUCT KERNEL MATRICES

By

Zhou Fan
Andrea Montanari

Technical Report No. 2015-14
July 2015

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

THE SPECTRAL NORM OF RANDOM INNER-PRODUCT KERNEL MATRICES

By

Zhou Fan
Andrea Montanari
Stanford University

Technical Report No. 2015-14
July 2015

**This research was supported in part by
National Science Foundation grants CCF 1319979 and DMS 1106627
and Air Force Office of Scientific Research grant FA9550-13-1-0036.**

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://statistics.stanford.edu

THE SPECTRAL NORM OF RANDOM INNER-PRODUCT KERNEL MATRICES

ZHOU FAN[1] AND ANDREA MONTANARI[1,2]

ABSTRACT. We study the spectra of $p \times p$ random matrices K with off-diagonal (i,j) entry equal to $n^{-1/2}k(X_i^T X_j/n^{1/2})$, where X_i's are the rows of a $p \times n$ matrix with i.i.d. entries and k is a scalar function. It is known that under mild conditions, as n and p increase proportionally, the empirical spectral measure of K converges to a deterministic limit μ. We prove that if k is a polynomial and the distribution of entries of X_i is symmetric and satisfies a general moment bound, then K is the sum of two components, the first with spectral norm converging to $\|\mu\|$ (the maximum absolute value of the support of μ) and the second a perturbation of rank at most two. In certain cases, including when k is an odd polynomial function, the perturbation is 0 and the spectral norm $\|K\|$ converges to $\|\mu\|$. If the entries of X_i are Gaussian, we also prove that $\|K\|$ converges to $\|\mu\|$ for a large class of odd non-polynomial functions k. In general, the perturbation may contribute spike eigenvalues to K outside of its limiting support, and we conjecture that they have deterministic limiting locations as predicted by a deformed GUE model. Our study of such matrices is motivated by the analysis of statistical thresholding procedures to estimate sparse covariance matrices from multivariate data, and our results imply an asymptotic approximation to the spectral norm error of such procedures when the population covariance is the identity.

1. INTRODUCTION

Let $X \in \mathbb{R}^{p \times n}$ be a random matrix with i.i.d. entries of zero mean and unit variance, and let $X_1^T, \ldots, X_p^T$ denote the rows of X. We study in this paper random matrices $K(X) \in \mathbb{R}^{p \times p}$ having entries $(K(X))_{ii'} = \frac{1}{\sqrt{n}} k\left(\frac{X_i^T X_{i'}}{\sqrt{n}}\right)$ for all $i \neq i'$ and $(K(X))_{ii} = 0$ for all i, for a function $k : \mathbb{R} \to \mathbb{R}$. Our main results pertain to the spectral properties of $K(X)$, specifically its spectral norm, in the asymptotic regime of random matrix theory where $n, p \to \infty$ proportionally with $\frac{p}{n} \to \gamma \in (0, \infty)$, and k is a fixed function independent of n and p. The study of such matrices $K(X)$ in this regime was initiated by Xiuyuan Cheng and Amit Singer in [9]. Following [9], we will call k a "kernel function" and $K(X)$ a "random inner-product kernel matrix".

Our study of this random matrix model is motivated by the following statistical application: Suppose $Y_1, \ldots, Y_n \in \mathbb{R}^p$ represent n i.i.d. observations of a random vector with mean zero and unknown covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$, and we wish to estimate Σ from these observations. If p is of comparable size to n, then the standard sample covariance matrix $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n Y_i Y_i^T$ is a poor estimator of Σ, and in general one cannot hope to estimate Σ with high accuracy. However, if it is known a priori that Σ is sufficiently sparse, i.e. most of its off-diagonal entries are zero, then it may still be possible to estimate Σ accurately under this assumption. A popular procedure for performing this estimation is to apply elementwise hard-thresholding to $\hat{\Sigma}$, i.e. to preserve those entries of $\hat{\Sigma}$ that are greater than τ in magnitude and to set the remaining entries to 0, for some threshold level $\tau := \tau(n,p)$ [4, 13]. Modifications of this procedure that apply continuous thresholding functions

[1]DEPARTMENT OF STATISTICS, STANFORD UNIVERSITY

[2]DEPARTMENT OF ELECTRICAL ENGINEERING, STANFORD UNIVERSITY

E-mail addresses: zhoufan@stanford.edu, montanari@stanford.edu.

ZF is supported by a Hertz Foundation Fellowship and an NDSEG Fellowship (DoD, Air Force Office of Scientific Research, 32 CFR 168a). AM is partially supported by NSF grants CCF-1319979 and DMS-1106627 and the AFOSR grant FA9550-13-1-0036.

elementwise to $\hat{\Sigma}$ have also been proposed and studied [25]. Such procedures have found application not only in problems where covariance estimation is the end goal, but also as subroutines of other statistical methods that require covariance estimation as an intermediate step, including procedures for sparse principal components analysis [18, 10] and sparse linear discriminant analysis [26].

If the true covariance matrix Σ is truly sparse in a suitable sense, and if the threshold level is set to $\tau = c\sqrt{\frac{\log p}{n}}$ for a sufficiently large constant $c > 0$, then it may be shown that the spectral norm error of such an estimator converges to zero as $n, p \to \infty$ [4, 13], and furthermore that the rate of convergence is minimax optimal over certain classes of sparse matrices [5]. However, these results do not provide an accurate estimate of how large one should expect this error to be for any specific choice of threshold level τ. To study this question, it is more natural to consider the asymptotic regime in which hard-thresholding is performed at the level $\tau = \frac{c}{\sqrt{n}}$ for a constant c, so that τ is of the same order as the typical "noise level" of the off-diagonal entries of $\hat{\Sigma}$. Then the entries of this thresholded covariance estimator are precisely given by $\frac{1}{\sqrt{n}} k\left(\frac{X_i^T X_i'}{\sqrt{n}}\right)$, where $X_i^T = (Y_{1i}, \ldots, Y_{ni})$, and $k(x)$ is a thresholding function (e.g. $k(x) = x\mathbb{1}\{|x| \geq c\}$ in the case of hard-thresholding). If the true covariance matrix Σ is $\mathrm{Id}_{p\times p}$, the $p \times p$ identity matrix, and Y_{ji} are in fact i.i.d., then we are led to the matrix model $K(X)$ studied in this paper. As the diagonal of $\hat{\Sigma}$ (thresholded or not) converges to $\mathrm{Id}_{p\times p}$ in spectral norm under weak conditions when $\Sigma = \mathrm{Id}_{p\times p}$, we define $K(X)$ with zero diagonal, so that the spectral norm $\|K(X)\|$ is asymptotically equivalent to the spectral norm error of the thresholded covariance estimator. One of our main results, Theorem 2.10 below, will imply that for an odd and continuously differentiable thresholding function k, if the data has standard Gaussian distribution and $\Sigma = \mathrm{Id}_{p\times p}$, then the spectral norm error of the thresholded covariance estimator converges almost surely to a positive deterministic constant in this asymptotic regime. Our theorem characterizes the dependence of this constant on the thresholding function k. For a given problem of size n and p and threshold level τ, this provides an asymptotic approximation to the error of the covariance thresholding estimator. We believe that many of the conditions of Theorem 2.10, such as the normality of the data and the continuous-differentiability of the threshold function k, may be relaxed.

Whether $\|K(X)\|$ converges is also a natural question to ask from the perspective of random matrix theory, and it was posed as an open question in [9]. For the identity kernel $k(x) = x$, $K(X)$ is equal to the sample covariance matrix $\frac{1}{n}XX^T$, excluding the diagonal. Under weak moment conditions on the entries x_{ij} of X, it is easily verified when $k(x) = x$ that $\|K(X) - (\frac{1}{n}XX^T - \mathrm{Id}_{p\times p})\| \to 0$ a.s. as $n, p \to \infty$. Hence, when $k(x) = x$, many asymptotic properties of the spectrum of $K(X)$, including the limit of its empirical spectral measure and the limit of its largest and smallest eigenvalues, match those of the matrix $\frac{1}{n}XX^T$ translated by -1. The asymptotic spectral behavior of $\frac{1}{n}XX^T$ in the large n and p limit is, by now, well-understood. The empirical spectral measure of $\frac{1}{n}XX^T$ converges weakly a.s. to a deterministic limit $\mu_{MP,\gamma}$ with compact support, known as the Marcenko-Pastur law [23]. Almost-sure convergence of the largest eigenvalue of $\frac{1}{n}XX^T$ to the upper endpoint of the support of $\mu_{MP,\gamma}$ was proven in [16] assuming certain moment conditions, and these conditions were later weakened in [34] to existence of the fourth moment of x_{ij}. Almost sure convergence of the smallest eigenvalue of $\frac{1}{n}XX^T$ to the lower endpoint of the support of $\mu_{MP,\gamma}$, when $\gamma < 1$, was shown in [1]. The fluctuations of the extremal eigenvalues of $\frac{1}{n}XX^T$ around their almost sure limits are also understood—for instance, assuming that $x_{ij} \sim \mathcal{N}(0, 1)$, it was shown in [17] that, after appropriate centering and rescaling, the distribution of the largest eigenvalue of $\frac{1}{n}XX^T$ converges weakly to the Tracy-Widom law of order 1, first introduced in [30] as the limiting distribution of the largest eigenvalue of the Gaussian Orthogonal Ensemble. This result has been extended to more general distributions of x_{ij} satisfying exponential decay conditions in [24].

For the case of a general kernel function k, the main theorem of [9] establishes that, if $x_{ij} \sim \mathcal{N}(0, 1)$, then the empirical spectral measure of $K(X)$ also converges weakly a.s. to a deterministic

limit, under certain mild assumptions on k. This limit distribution, which we denote as $\mu_{a,\nu,\gamma}$ (c.f. Definition 2.3 below), depends on k through its orthogonal decomposition in the Hermite polynomial basis of $L_2(q(x)dx)$, where $q(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ is the standard Gaussian density. When $k(x) = x$, $\mu_{a,\nu,\gamma}$ coincides with the Marcenko-Pastur law $\mu_{MP,\gamma}$ translated by -1. Quite remarkably, if k has no linear component in its Hermite polynomial decomposition, then the limiting spectral distribution $\mu_{a,\nu,\gamma}$ is an appropriately scaled version of Wigner's semicircle law. In the general setting, a characterization of the limiting measure $\mu_{a,\nu,\gamma}$ was given in [9] in terms of an implicit equation in its Stieltjes transform, restated as eq. (1) below. This result was shown to hold for more general distributions of x_{ij} having moments of all orders in Theorem 3 of [11].

We observe that the limiting spectral distribution $\mu_{a,\nu,\gamma}$ of $K(X)$ is in fact the additive free convolution, as defined by Dan Voiculescu in [32], of a scaled semicircle law and a scaled and translated Marcenko-Pastur law. As such, it is also the limiting spectral distribution of a random matrix $W+V$, for $W \in \mathbb{C}^{p\times p}$ a real symmetric or complex Hermitian Wigner matrix and $V \in \mathbb{R}^{p\times p}$ a deterministic diagonal matrix whose empirical spectral measure converges to this scaled and translated Marcenko-Pastur law [33, 12]. Such matrices have been referred to as "deformed Wigner matrices" or "Wigner matrices with external source" in the random matrix theory literature, and they have been the subject of much recent study. For instance, if W is distributed as the Gaussian Unitary Ensemble and V does not have eigenvalues outside of the support of its limiting spectrum, then the results of [6] and [22] imply that in the limit of large n and p, no eigenvalues of $W + V$ fall outside of $\mathrm{supp}(\mu_{a,\nu,\gamma}) + (-\varepsilon, \varepsilon)$ a.s. for any $\varepsilon > 0$, and in particular, the largest and smallest eigenvalues of $W + V$ converge to the upper and lower boundaries of $\mathrm{supp}(\mu_{a,\nu,\gamma})$. More generally, if V has a finite number of fixed "spike" eigenvalues outside of its limiting support, then $W + V$ may have corresponding spike eigenvalues outside of $\mathrm{supp}(\mu_{a,\nu,\gamma})$ as well, and the conditions for existence and the limiting locations of these spike eigenvalues of $W + V$ are fully characterized in [6]. Under various assumptions on W and V, more detailed results regarding the fluctuations of the spike eigenvalues of $W + V$ and the eigenvalues at the edges of $\mathrm{supp}(\mu_{a,\nu,\gamma})$ have also been obtained, see e.g. [27, 7, 20].

In comparison, the spectral norm and largest and smallest eigenvalues of the random inner-product kernel matrix $K(X)$ are not well-understood when the kernel function k is nonlinear. In previous work, Lemma 4.2 of [9] showed that $\mathbb{E}\|K(X)\| \leq O_d(n^{1/4})$ when the kernel function $k(x) := k_n(x)$ is the degree-d orthogonal polynomial with respect to the distribution of $\frac{X_i^T X_{i'}}{\sqrt{n}}$, and the authors conjectured that the true size of $\mathbb{E}\|K(X)\|$ in this case should be $O_d(1)$. Proposition 6.2 of [10] used Gaussian concentration of measure results and a covering argument to show that, when x_{ij} have Gaussian distribution, $\|K(X)\| \leq O_\tau(1)$ with high probability when $k(x) = \mathrm{sgn}(x)(|x|-\tau)_+$ is the soft-thresholding function at level $\tau > 0$. We are not aware of existing results that establish whether $\|K(X)\|$, $\lambda_{\max}(K(X))$, and $\lambda_{\min}(K(X))$ converge to the respective quantities $\sup\{|x| : x \in \mathrm{supp}(\mu_{a,\nu,\gamma})\}$, $\sup\{x : x \in \mathrm{supp}(\mu_{a,\nu,\gamma})\}$, and $\inf\{x : x \in \mathrm{supp}(\mu_{a,\nu,\gamma})\}$ for any nonlinear kernel function k.

In this paper, we provide an analysis of the spectral norm $\|K(X)\|$ for the case where k is a polynomial function and the distribution of x_{ij} is symmetric and satisfies Assumption 2.1 below. Our results imply (Corollary 2.7) that $\|K(X)\|$ converges a.s. to $\|\mu_{a,\nu,\gamma}\| := \sup\{|x| : x \in \mathrm{supp}(\mu_{a,\nu,\gamma})\}$ either when $x_{ij} \sim \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$ or when the Hermite polynomial decomposition of k has no degree-2 component. The latter condition holds, in particular, if k is an odd polynomial function. More generally (Theorem 2.6), we exhibit a decomposition $K(X) = \tilde{K}(X) + \tilde{R}(X)$, where the empirical spectral distribution of $\tilde{K}(X)$ converges weakly a.s. to $\mu_{a,\nu,\gamma}$, $\|\tilde{K}(X)\|$ converges a.s. to $\|\mu_{a,\nu,\gamma}\|$, and $\tilde{R}(X)$ is a rank-two perturbation matrix whose two nonzero eigenvalues converge to fixed quantities. In the limit of large n and p, $\tilde{R}(X)$ may contribute "spike" eigenvalues to $K(X)$ that fall outside of $\mathrm{supp}(\mu_{a,\nu,\gamma})$ and in particular may be larger in magnitude than $\|\mu_{a,\nu,\gamma}\|$, and we conjecture that

these spike eigenvalues have deterministic limiting locations that match those of a deformed Wigner model as characterized in [6].

We believe that our main result may be extended to certain classes of non-polynomial functions k via polynomial approximation arguments, although the details are non-trivial. As one instance of such an extension, we show that if k is an odd and continuously-differentiable function whose derivative $k'(x)$ grows at most exponentially in x, then $\|K(X)\|$ converges to $\|\mu_{a,\nu,\gamma}\|$ when x_{ij} have Gaussian distribution (Theorem 2.10). Such an extension is important for the applicability of our results to the statistical covariance estimation application previously discussed, and further relaxations of the conditions imposed on k and the distribution of x_{ij} are an interesting avenue for future work.

Let us remark that in order for the empirical spectrum of a kernel random matrix to converge to a deterministic limit, the entries of the matrix must be scaled appropriately as n and p increase. When k is nonlinear, scaling inside and outside of the kernel function k are not interchangeable and can lead to very different asymptotic behaviors. When x_{ij} are independent with unit variance, $X_i^T X_{i'}$ is of typical size $O(n)$ for $i = i'$ and $O(\sqrt{n})$ for $i \neq i'$. In [14], Nourredine El Karoui first explored the spectral behavior of kernel random matrices in the large n and p limit under the scaling $(\bar{K}(X))_{ii'} = k\left(\frac{X_i^T X_{i'}}{n}\right)$ for a fixed function k. Under this scaling, k is applied to diagonal entries of size $O(1)$ and to off-diagonal entries of size $O\left(\frac{1}{\sqrt{n}}\right)$. Theorem 2.1 of [14] implies that if k is locally smooth at 0 and 1, then $\|\bar{K}(X) - M(X)\| \to 0$ for the simpler matrix $M(X) = \left(k(0) + \frac{k''(0)}{2n}\right)\mathbf{1}_p\mathbf{1}_p^T + k'(0)\frac{XX^T}{n} + (k(1) - k(0) - k'(0))\,\mathrm{Id}_{p\times p}$. In particular, the off-diagonal entries of $M(X)$ depend on k only through its Taylor expansion at 0. As a consequence of this result, the limiting spectral distributions of $\bar{K}(X)$ and $M(X)$ are identical, and it is given by shifting and rescaling the Marcenko-Pastur law. Furthermore, the largest and smallest eigenvalues of $\bar{K}(X)$ and $M(X)$ have the same almost sure limits. It was concluded in [14] that the asymptotic spectral properties of $\bar{K}(X)$ are "essentially linear". In contrast, we remove the matrix diagonal and apply the scaling $(K(X))_{ii'} = \frac{1}{\sqrt{n}}k\left(\frac{X_i^T X_{i'}}{\sqrt{n}}\right)$ to the off-diagonal entries, again for a fixed function k. This scaling applies k to off-diagonal entries of typical size $O(1)$ and yields the more interesting "nonlinear" behavior discovered in [9]. Consideration of this scaling is strongly motivated by the covariance thresholding application previously discussed.

A formal statement of our definitions, assumptions, theorems, and conjectures, as well as a high-level outline of the proof, are provided in Section 2. The proof of our main result regarding polynomial kernel functions k is given in Sections 3–5, with details deferred to two appendices. Finally, the proof of our extension to odd and continuously-differentiable kernel functions in the Gaussian case is given in Section 6.

2. Main theorems and discussion

2.1. **Background and statement of results.** Let $X = (x_{ij} : 1 \le i \le p, 1 \le j \le n) \in \mathbb{R}^{p\times n}$ be a matrix whose entries x_{ij} are a collection of independent and identically distributed real random variables, satisfying the following conditions:

Assumption 2.1. (1) $\mathbb{E}[x_{ij}] = 0$ *and* $\mathbb{E}[x_{ij}^2] = 1$.
(2) $\mathbb{E}[|x_{ij}|^k] \le k^{\alpha k}$ *for all* $k \ge 2$ *and some* $\alpha > 0$.
(3) *The distribution of* x_{ij} *is symmetric, i.e.* $x_{ij} \overset{L}{=} -x_{ij}$.

Let $X_i^T = (x_{ij} : 1 \le j \le n) \in \mathbb{R}^n$ denote the i^{th} row of X.

Definition 2.2. *For a* **kernel function** $k : \mathbb{R} \to \mathbb{R}$, *and* $X \in \mathbb{R}^{p\times n}$, *the* **random inner-product kernel matrix** *associated to* k *and* X *is given by* $K_{n,p}(X) = (k_{ii'} : 1 \le i, i' \le p) \in \mathbb{R}^{p\times p}$ *with*

entries

$$k_{ii'} = \begin{cases} \frac{1}{\sqrt{n}} k\left(\frac{X_i^T X_{i'}}{\sqrt{n}}\right), & i \neq i' \\ 0, & i = i'. \end{cases}$$

Throughout, we will use capital variables to denote matrices and vectors, and lowercase variables to denote (random or deterministic) scalars. We will use $i, i', i_1, i_2, \ldots$ to denote indices in $\{1, \ldots, p\}$ and $j, j', j_1, j_2, \ldots$ to denote indices in $\{1, \ldots, n\}$. The notation $K_{n,p}$ refers to the a matrix whose definition depends on n and p; $k_{ii'}$ refers to the (i, i') entry of $K_{n,p}$, and for convenience we suppress the dependence of $k_{ii'}$ on n, p.

Assumption 2.1 specifies conditions on the distribution of x_{ij}. The moment condition in part (2) of Assumption 2.1 was also assumed in the analysis of the spectral norm of standard sample covariance matrices in [16], and in particular, it is satisfied by any sub-Gaussian or sub-exponential random variable (c.f. Definitions 5.7 and 5.13 of [31]). It is possible to weaken this assumption using truncation arguments, and we have not made an attempt to do so here. Part (3) of Assumption 2.1, that the distribution of x_{ij} is symmetric, is required for our subsequent combinatorial arguments, although it is probably not a necessary condition for our main results to hold.

The study of matrices $K_{n,p}(X)$ in Definition 2.2 in the asymptotic regime of large n and p was initiated by Xiuyuan Cheng and Amit Singer in [9], in which it was shown that the empirical spectral distribution of $K_{n,p}(X)$ converges weakly a.s. to a deterministic limit. To describe this limit, recall that for a probability measure μ on $\mathbb{R}$, its Stieltjes transform is the function $m : \mathbb{C}^+ \to \mathbb{C}^+$ given by

$$m(z) = \int \frac{\mu(d\lambda)}{\lambda - z},$$

where $\mathbb{C}^+ = \{z \in \mathbb{C} : \operatorname{Im} z > 0\}$ is the upper-half complex plane. The measure μ is uniquely determined by its Stieltjes transform and may be recovered from the inversion formula

$$\mu([a, b]) = \lim_{\varepsilon \to 0} \frac{1}{\pi} \int_a^b \operatorname{Im} m(\lambda + i\varepsilon) d\lambda$$

for any a, b that are continuity points of μ. (See Theorem B.8 of [2].)

Definition 2.3. *For $\nu, \gamma > 0$ and $a \in [-\sqrt{\nu}, \sqrt{\nu}]$, let $m(z) := m_{a,\nu,\gamma}(z)$ be the unique solution to the equation*

$$-\frac{1}{m(z)} = z + a\left(1 - \frac{1}{1 + a\gamma m(z)}\right) + \gamma(\nu - a^2)m(z), \qquad z \in \mathbb{C}^+ \tag{1}$$

with $m(z) \in \mathbb{C}^+$. Let $\mu_{a,\nu,\gamma}$ be the measure on $\mathbb{R}$ having Stieltjes transform $m(z)$, let $\operatorname{supp}(\mu_{a,\nu,\gamma})$ denote its support, and let $\|\mu_{a,\nu,\gamma}\| = \sup\{|x| : x \in \operatorname{supp}(\mu_{a,\nu,\gamma})\}$.

It was shown in [9] that $m(z)$ is well-defined and that $\mu_{a,\nu,\gamma}$ is a probability measure with continuous density and compact support. (Note that in our notation, n and p are reversed from their definitions in [9], and $\gamma = \lim_{n,p\to\infty} \frac{p}{n}$ corresponds to $\frac{1}{\gamma}$ in [9].) The following result establishes weak convergence of the empirical spectral measure of $K_{n,p}(X)$ to $\mu_{a,\nu,\gamma}$; it is implied by Theorem 3 of [11] and Remark 3.2 and Lemma C.2 of [9]. The result was first shown in [9] in the case where $x_{ij} \sim \mathcal{N}(0, 1)$ and was extended to more general distributions of x_{ij} in [11] via Lindeberg's swapping argument.

Definition 2.4. *Let $q(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ be the standard Gaussian density. Let $\{h_d\}_{d=0}^{\infty}$ be the Hermite polynomials orthonormal with respect to $L^2(q(x)dx)$, i.e. h_d is of degree d and, for $\xi \sim \mathcal{N}(0, 1)$, $\mathbb{E}[h_d(\xi)h'_d(\xi)] = 1$ if $d = d'$ and 0 if $d \neq d'$.*

Theorem 2.5 ([9, 11])**.** *Suppose $k : \mathbb{R} \to \mathbb{R}$ satisfies $\mathbb{E}[k(\xi)] = 0$ and $\mathbb{E}[k(\xi)^2] < \infty$ for $\xi \sim \mathcal{N}(0, 1)$. Let $k(x) = \sum_{d=1}^{\infty} a_d h_d(x)$ be the expansion of k in the Hermite polynomial basis, where convergence*

of the sum is in the sense of $L^2(q(x)dx)$, and let $a = a_1 = \mathbb{E}[\xi k(\xi)]$ and $\nu = \sum_{d=1}^{\infty} a_d^2 = \mathbb{E}[k(\xi)^2]$. Let $X \in \mathbb{R}^{p\times n}$ be a random matrix with i.i.d. entries having finite moments of all orders, and let X_i^T denote the i^{th} row of X. Suppose that $\left|\mathbb{E}\left[k(\xi)^2\right] - \mathbb{E}\left[k\left(\frac{X_1^T X_2}{\sqrt{n}}\right)^2\right]\right| \to 0$ as $n \to \infty$, let $K_{n,p}(X)$ be as in Definition 2.2 with kernel function k, and let $\lambda_1, \ldots, \lambda_p$ be the eigenvalues of $K_{n,p}(X)$. Then, as $n, p \to \infty$ with $\frac{p}{n} \to \gamma \in (0, \infty)$,

$$\frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_i} \Rightarrow \mu_{a,\nu,\gamma},$$

where $\frac{1}{p}\sum_{i=1}^{p}\delta_{\lambda_i}$ is the empirical spectral measure of $K_{n,p}(X)$, $\mu_{a,\nu,\gamma}$ is the measure in Definition 2.3, and the convergence holds weakly a.s.

This weak convergence result does not necessarily imply convergence of $\|K_{n,p}(X)\|$ to $\|\mu_{a,\nu,\gamma}\|$, as it only ensures that there are at most $o(n)$ eigenvalues of $K_{n,p}(X)$ falling outside of the support of $\mu_{a,\nu,\gamma}$. The following theorem, which is the first main result of this paper, examines this question of convergence of spectral norm in the case where the kernel $k(x)$ is a polynomial function.

Theorem 2.6. *Suppose $k(x) = \sum_{a=1}^{D} a_d h_d(x)$ where h_d is the Hermite polynomial of degree d as in Definition 2.4 (so $k(x)$ is a polynomial function with $\mathbb{E}[k(\xi)] = 0$ for $\xi \sim \mathcal{N}(0,1)$). Let $a = a_1 = \mathbb{E}[\xi k(\xi)]$ and $\nu = a_1^2 + \ldots + a_D^2 = \mathbb{E}[k(\xi)^2]$. Let X satisfy Assumption 2.1, let $K_{n,p}(X)$ be as in Definition 2.2 with kernel function k, and suppose $n, p \to \infty$ with $\frac{p}{n} \to \gamma$. Then $K_{n,p}(X) = \tilde{K}_{n,p}(X) + \tilde{R}_{n,p}(X)$, where*

1. *$\lim_{n,p\to\infty} \|\tilde{K}_{n,p}(X)\| = \|\mu_{a,\nu,\gamma}\|$ a.s.*
2. *$\tilde{R}_{n,p}(X)$ is of rank at most two. Specifically, letting $V_{n,p}(X) = (v_i : 1 \le i \le p) \in \mathbb{R}^p$ be the vector with entries $v_i = \sum_{j=1}^{n} \frac{x_{ij}^2 - 1}{\sqrt{n}}$ and denoting $\mathbf{1}_p = (1, \ldots, 1) \in \mathbb{R}^p$,*

$$\tilde{R}_{n,p}(X) = \frac{a_2}{n\sqrt{2}}(V_{n,p}(X)\mathbf{1}_p^T + \mathbf{1}_p V_{n,p}(X)^T).$$

Note that in this theorem, $\tilde{K}_{n,p}(X)$ also has $\mu_{a,\nu,\gamma}$ as its limiting spectral measure, since this limit is unaffected by perturbations of finite rank. (See Theorem A.43 of [2].) Hence this theorem characterizes $K_{n,p}(X)$ as the sum of two components, the first of which has spectral norm converging to $\|\mu_{a,\nu,\gamma}\|$ as expected, and the second of which might contribute additional spike eigenvalues to $K_{n,p}(X)$ that may be larger in magnitude than $\|\mu_{a,\nu,\gamma}\|$. This theorem has the following immediate corollaries.

Corollary 2.7. *Under the assumptions of Theorem 2.6, if $a_2 = 0$ or if $x_{ij} \sim \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$, then $\lim_{n,p\to\infty} \|K_{n,p}(X)\| = \|\mu_{a,\nu,\gamma}\|$ almost surely. In particular, if k is an odd polynomial function, i.e. $k(-x) = -k(x)$, then $\|K_{n,p}(X)\| = \|\mu_{a,\nu,\gamma}\|$ almost surely.*

Proof. If $a_2 = 0$ or $x_{ij} \sim \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$, then $\tilde{R}_{n,p}(X) = 0$, so the result follows from Theorem 2.6. If k is an odd function, then $a_2 = 0$. □

Corollary 2.8. *Under the assumptions of Theorem 2.6, if $a_2 \neq 0$ and x_{ij} is not distributed as $\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$, then the two non-zero eigenvalues of $\tilde{R}_{n,p}(X)$ converge a.s. to $\pm a_2\gamma\sqrt{\frac{\mathbb{E}x_{ij}^4 - 1}{2}}$.*

Proof. Letting $V := V_{n,p}(X) = \{v_i : 1 \le i \le p\}$ be as in Theorem 2.6, we may compute $\operatorname{Tr}\tilde{R}_{n,p}(X) = \frac{a_2\sqrt{2}}{n}\sum_{i=1}^{p} v_i$ and $\operatorname{Tr}\tilde{R}_{n,p}(X)^2 = \frac{a_2^2}{n^2}\left(\left(\sum_{i=1}^{p} v_i\right)^2 + p\|V\|_2^2\right)$. If λ_1 and λ_2 are the two non-zero eigenvalues of $\tilde{R}_{n,p}(X)$, then this implies $\lambda_1 + \lambda_2 = \frac{a_2\sqrt{2}}{n}\sum_{i=1}^{p} v_i$ and $\lambda_1\lambda_2 =$

$\frac{(\lambda_1+\lambda_2)^2-\lambda_1^2-\lambda_2^2}{2} = \frac{a_2^2}{2n^2}\left(\left(\sum_{i=1}^p v_i\right)^2 - p\|V\|_2^2\right)$, so λ_1 and λ_2 are the roots of the equation

$$\lambda^2 - \left(\frac{a_2\sqrt{2}}{n}\sum_{i=1}^p v_i\right)\lambda + \frac{a_2^2}{2n^2}\left(\left(\sum_{i=1}^p v_i\right)^2 - p\|V\|_2^2\right) = 0.$$

By the law of large numbers, $\lim_{n,p\to\infty}\frac{1}{n}\sum_{i=1}^p v_i = 0$ and $\lim_{n,p\to\infty}\frac{p\|V\|_2^2}{n^2} \to \gamma^2(\mathbb{E}x_{ij}^4 - 1)$ almost surely. Since the roots of a polynomial are continuous in its coefficients, the result follows. □

Corollary 2.9. *Under the assumptions of Theorem 2.6,*

$$\limsup_{n,p\to\infty}\|K_{n,p}(X)\| \le \|\mu_{a,\nu,\gamma}\| + |a_2|\gamma\sqrt{\frac{\mathbb{E}x_{ij}^4 - 1}{2}} < \infty$$

almost surely.

Proof. As $\|K_{n,p}(X)\| \le \|\tilde{K}_{n,p}(X)\| + \|\tilde{R}_{n,p}(X)\|$, this follows from Theorem 2.6 and Corollary 2.8. □

Corollary 2.7 implies that if $k(x)$ is an odd polynomial function, then $\|K_{n,p}(X)\| \to \|\mu_{a,\nu,\gamma}\|$. In the case where x_{ij} have Gaussian distribution, we extend this conclusion to more general odd kernel functions in our second main result.

Theorem 2.10. *Suppose $k:\mathbb{R}\to\mathbb{R}$ is an odd function, i.e. $k(-x) = -k(x)$, and that it is continuously differentiable with $\limsup_{|x|\to\infty}\frac{\log|k'(x)|}{|x|} < \infty$. Let $k(x) = \sum_{d=1}^\infty a_d h_d(x)$ be the expansion of k in the Hermite polynomial basis of Definition 2.4, where convergence of the sum is in the sense of $L^2(q(x)dx)$, and let $a = a_1 = \mathbb{E}[\xi k(\xi)]$ and $\nu = \sum_{d=1}^\infty a_d^2 = \mathbb{E}[k(\xi)^2]$ for $\xi\sim\mathcal{N}(0,1)$. Let $X\in\mathbb{R}^{p\times n}$ have entries $x_{ij}\overset{iid}{\sim}\mathcal{N}(0,1)$, let $K_{n,p}(X)$ be as in Definition 2.2 with kernel function k, and let $n,p\to\infty$ with $\frac{p}{n}\to\gamma$. Then, almost surely,*

$$\lim_{n,p\to\infty}\|K_{n,p}(X)\| = \|\mu_{a,\nu,\gamma}\|.$$

Note that, by the assumptions of Theorem 2.10, $|k(x)| \le Ce^{C|x|}$ for some constant $C > 0$ and all $x\in\mathbb{R}$, so $\nu = \mathbb{E}[k(\xi)^2] < \infty$ and the Hermite polynomial decomposition of k is well-defined.

2.2. **Discussion and conjectures.** We observe in the following proposition that the limiting measure $\mu_{a,\nu,\gamma}$ of Definition 2.3 is the additive free convolution [32] of a semicircle law and the Marcenko-Pastur law translated by -1 and scaled by a. Recall that for a probability measure μ with Stieltjes transform $m(z)$, there exist $\eta, M > 0$ such that $z\mapsto -\frac{1}{m(z)}$ is injective on $\{z\in\mathbb{C}^+ : \operatorname{Im} z > \eta\operatorname{Re} z, |z| > M\}$. Denoting the inverse of $z\mapsto -\frac{1}{m(z)}$ on this domain as $S(z)$, the Voiculescu R-transform of μ is given by the function $R(z) = S(\frac{1}{z}) - \frac{1}{z}$. For two probability measures μ and ν on $\mathbb{R}$ with R-transforms R_μ and R_ν, their free additive convolution $\mu\boxplus\nu$ is the probability measure whose R-transform is given by $R_{\mu\boxplus\nu}(z) = R_\mu(z) + R_\nu(z)$ [32, 21, 3].

Proposition 2.11. *Let μ_{sc} be the semicircle law with density*

$$\mu_{sc}(dx) = \sqrt{4\gamma(\nu-a^2) - x^2}\,\mathbb{1}_{\left[-2\sqrt{\gamma(\nu-a^2)},2\sqrt{\gamma(\nu-a^2)}\right]}(x)dx.$$

Let $\mu_{MP,shift}$ be the Marcenko-Pastur law with parameter γ, translated by -1 and scaled by a. (If $\gamma\le 1$ and $a > 0$, then $\mu_{MP,shift}$ has density

$$\mu_{MP,shift}(dx) = \frac{1}{2\pi}\frac{\sqrt{\left(\gamma+2\sqrt{\gamma}-\frac{x}{a}\right)\left(\frac{x}{a}-\gamma+2\sqrt{\gamma}\right)}}{\gamma(x+a)}\mathbb{1}_{[a\gamma-2a\sqrt{\gamma},a\gamma+2a\sqrt{\gamma}]}(x)dx.$$

If $\gamma > 1$ and $a > 0$, then $\mu_{MP,shift}$ is a mixture $\left(1 - \frac{1}{\gamma}\right)\delta_0 + \frac{1}{\gamma}\tilde{\mu}_{MP,shift}$ where $\tilde{\mu}_{MP,shift}$ has the above density. If $a = 0$, then $\mu_{MP,shift} = \delta_0$.) Then $\mu_{a,\nu,\gamma}$ given in Definition 2.3 is equal to the free additive convolution $\mu_{sc} \boxplus \mu_{MP,shift}$.

Proof. The semicircle law μ_{sc} has Stieltjes transform

$$m_{sc}(z) = -\frac{1}{2\gamma(\nu - a^2)}\left(z - \sqrt{z^2 - 4\gamma(\nu - a^2)}\right),$$

and when $a > 0$, $\mu_{MP,shift}$ has Stieltjes transform

$$m_{MP,shift}(z) = \frac{-\gamma - \frac{z}{a} + \sqrt{\left(\frac{z}{a} - \gamma\right)^2 - 4\gamma}}{2\gamma(z + a)}.$$

Here, the branch of the square-root function is chosen to have positive imaginary part. (See Lemmas 2.11 and 3.11 of [2].) The first equation implies $-\frac{1}{m_{sc}(z)} = z + \gamma(\nu - a^2)m_{sc}(z)$. When $a = 0$, this is precisely eq. (1) defining $\mu_{a,\nu,\gamma}$, so $\mu_{sc} = \mu_{a,\nu,\gamma}$. When $a > 0$, the second equation implies $-\frac{1}{m_{MP,shift}(z)} = z + a\left(1 - \frac{1}{1+a\gamma m_{MP,shift}(z)}\right)$, and the R-transforms of μ_{sc} and $\mu_{MP,shift}$ are given by $R_{sc}(z) = \gamma(\nu - a^2)z$ and $R_{MP,shift}(z) = -a\left(1 - \frac{1}{1-a\gamma z}\right)$, respectively. Then the R-transform of $\mu_{sc} \boxplus \mu_{MP,shift}$ is given by $R_{\mu_{sc}\boxplus\mu_{MP,shift}}(z) = -a\left(1 - \frac{1}{1-a\gamma z}\right) + \gamma(\nu - a^2)z$, so the Stieltjes transform $m_{\mu_{sc}\boxplus\mu_{MP,shift}}(z)$ satisfies

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

at least on an open sub-domain of $\mathbb{C}^+$. This agrees with eq. (2.3) defining the Stieltjes transform of $\mu_{a,\nu,\gamma}$, and as the Stieltjes transform of any measure is analytic on $\mathbb{C}^+$, these functions must agree on all of $\mathbb{C}^+$. Then $\mu_{a,\nu,\gamma} = \mu_{sc} \boxplus \mu_{MP,shift}$. □

In particular, if $a = 0$, then the measure $\mu_{a,\nu,\gamma}$ is the semicircle law μ_{sc}, and if $a^2 = \nu$, then it is the translated and scaled Marcenko-Pastur law $\mu_{MP,shift}$. (See also Remarks 3.6 and 3.7 of [9].) By Remark 2.2 of [6], when $\gamma \leq 1$ or $a = 0$, the support of $\mu_{a,\nu,\gamma}$ must be a single interval, and when $\gamma > 1$ and $a > 0$, it must either be a single interval or the union of two disjoint intervals. We observe that when $a = 0$, the measure $\mu_{a,\nu,\gamma}$ is symmetric, and hence Theorem 2.6 implies the almost sure convergence of both the largest and smallest eigenvalues of $\tilde{K}_{n,p}(X)$ to $\pm\|\mu_{a,\nu,\gamma}\|$. In general, $\mu_{a,\nu,\gamma}$ is not symmetric, and Theorem 2.6 provides information on only one of these two eigenvalues. Furthermore, when $\mu_{a,\nu,\gamma}$ has two disjoint intervals of support, Theorem 2.6 does not describe whether $\tilde{K}_{n,p}(X)$ has eigenvalues in between these two intervals of support. We conjecture that, in fact, $\tilde{K}_{n,p}(X)$ has no eigenvalues outside of the limiting support, in the following sense.

Conjecture 2.12. Under the assumptions of Theorem 2.6, for any $\varepsilon > 0$,

$$\lim_{n,p\to\infty} \mathbb{P}[\exists i \in \{1, \ldots, p\} : \operatorname{dist}(\lambda_i(\tilde{K}_{n,p}(X)), \operatorname{supp}(\mu_{a,\nu,\gamma})) > \varepsilon] = 0,$$

where $\operatorname{dist}(x, C) = \inf\{|x - y| : y \in C\}$.

Turning to the rank-two matrix $\tilde{R}_{n,p}(X)$ of Theorem 2.6, we note that $\tilde{R}_{n,p}(X)$ may contribute "spike" eigenvalues to $K_{n,p}(X)$ that fall outside of the support of the limiting spectral measure $\mu_{a,\nu,\gamma}$. The entries of $\tilde{R}_{n,p}(X)$ and $\tilde{K}_{n,p}(X)$ are dependent, but let us momentarily consider the simpler matrix $W + V$, where W is a Wigner Hermitian matrix with limiting empirical spectral measure μ_{sc} and V is a deterministic diagonal matrix with limiting empirical spectral measure $\mu_{MP,shift}$, for μ_{sc} and $\mu_{MP,shift}$ as defined in Proposition 2.11. Then by [33], the empirical spectral measure of $W + V$ converges to $\mu_{a,\nu,\gamma}$. Suppose, in addition, that V has at most two "spike"

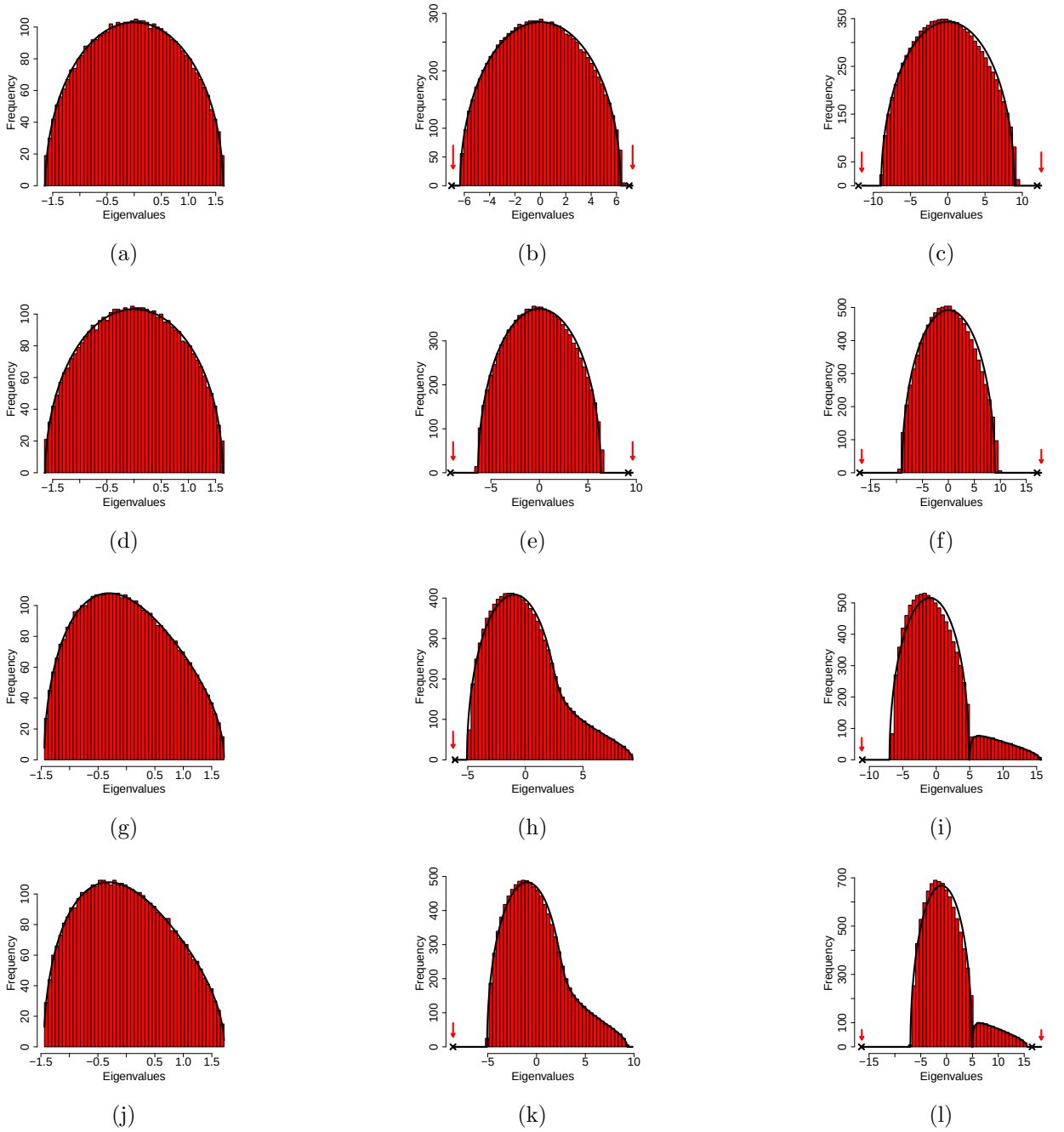

FIGURE 1. Simulation results are shown for the empirical spectra and extremal eigenvalues of matrices $K_{n,p}(X)$, for various kernel functions, distributions of the matrix entries x_{ij}, and parameters a, ν, and γ. The histogram of eigenvalues of $K_{n,p}(X)$ is shown in red, with the (scaled) density function of the limiting empirical spectral measure $\mu_{a,\nu,\gamma}$ shown in black. In settings where Conjecture 2.13 predicts the presense of spike eigenvalues of $K_{n,p}(X)$, the theoretical predictions for the locations of these spikes according to Conjecture 2.13 are shown as black crosses, and the locations of the corresponding largest and/or smallest empirically observed eigenvalues of $K_{n,p}(X)$ are indicated by red arrows. Panels (a)–(f) correspond to the kernel function $k(x) = h_2(x) + h_3(x)$. Panels (g)–(l) correspond to the kernel function $k(x) = 0.9h_1(x) + h_2(x)$. Panels (a), (d), (g), and (j) correspond to $p = 4000$, $n = 12000$, and $\gamma = \frac{1}{3}$. Panels (b), (e), (h), and (k) correspond to $p = 10000$, $n = 2000$, and $\gamma = 5$. Panels (c), (f), (i), and (l) correspond to $p = 10000$, $n = 1000$, and $\gamma = 10$. Panels (a)–(c) and (g)–(i) take $x_{ij} \sim \mathcal{N}(0, 1)$, so $\mathbb{E}[x_{ij}^4] = 3$, and panels (d)–(f) and (j)–(l) take $x_{ij} \sim \text{Laplace}(0, \frac{1}{\sqrt{2}})$, so $\mathbb{E}[x_{ij}^4] = 6$.

eigenvalues outside of $\mathrm{supp}(\mu_{MP,shift})$, and these eigenvalues are fixed and equal to $\pm a_2\gamma\sqrt{\frac{\mathbb{E}x_{ij}^4-1}{2}}$, as given in Corollary 2.8. Then Theorem 8.1 of [6] implies that $W+V$ has at most two "spike" eigenvalues outside of $\mathrm{supp}(\mu_{a,\nu,\gamma})$ and gives a precise characterization of when such spikes occur and the locations at which they appear. We conjecture that this characterization, summarized below, holds also for the spike eigenvalues of $K_{n,p}(X)$, even though $\tilde{K}_{n,p}(X)$ is not a deformed Wigner matrix and $\tilde{R}_{n,p}(X)$ is not independent of $\tilde{K}_{n,p}(X)$.

Conjecture 2.13. Consider the setup of Theorem 2.6, and suppose $a^2 < \nu$. Let $S = \mathrm{supp}(\mu_{MP,\gamma})$ and define $H : \mathbb{R} \setminus S \to \mathbb{R}$ by $H(z) = z - \gamma(\nu - a^2)m_{MP,\gamma}(z)$, where $\mu_{MP,\gamma}$ is the scaled and translated Marcenko-Pastur law as in Proposition 2.11 and $m_{MP,\gamma}(z)$ is its Stieltjes transform. Let $\lambda_1, \lambda_2 = \pm a_2\gamma\sqrt{\frac{\mathbb{E}x_{ij}^4-1}{2}}$. If $\lambda_1 \notin S$ and $H'(\lambda_1) > 0$, then $H(\lambda_1) \notin \mathrm{supp}(\mu_{a,\nu,\gamma})$ and there is one eigenvalue of $K_{n,p}(X)$ that converges a.s. to $H(\lambda_1)$. Similarly, if $\lambda_2 \notin S$ and $H'(\lambda_2) > 0$, then $H(\lambda_2) \notin \mathrm{supp}(\mu_{a,\nu,\gamma})$ and there is one eigenvalue of $K_{n,p}(X)$ that converges a.s. to $H(\lambda_2)$. The remaining eigenvalues of $K_{n,p}(X)$ are, for any $\varepsilon > 0$, within an ε-neighborhood of $\mathrm{supp}(\mu_{a,\nu,\gamma})$, in the sense of Conjecture 2.12.

Figure 1 depicts simulation results of the empirical spectrum of $K_{n,p}(X)$ for the kernel functions $k(x) = h_2(x) + h_3(x)$ and $k(x) = 0.9h_1(x) + h_2(x)$, various settings of $\gamma = \frac{p}{n}$, and Gaussian and Laplace distributions for x_{ij}. For the parameter combinations for which Conjecture 2.13 predicts the presence of spike eigenvalues in the limiting spectrum of $K_{n,p}(X)$, we in fact observe such spike eigenvalues in simulation. The predicted spike locations are shown together with the empirically observed locations, and there is close agreement in all cases.

2.3. **Outline of proof.** In the remainder of the paper, we prove Theorems 2.6 and 2.10. Our proof of Theorem 2.6 follows three high-level steps:

(1) If $z_1, \ldots, z_n$ are i.i.d. random variables with mean zero, variance one, and zero third moment, we show that

$$\sqrt{d!}h_d\left(\frac{\sum_{i=1}^n z_i}{\sqrt{n}}\right) \approx \sqrt{\frac{1}{n^d}} \sum_{\substack{j_1,\ldots,j_d=1 \\ j_1\neq j_2\neq\ldots\neq j_d}}^{n} \prod_{i=1}^{d} z_{j_i}, \tag{2}$$

where h_d is the degree-d orthonormal Hermite polynomial as in Definition 2.4. We note that $\sqrt{d!}h_d$ on the left side of eq. (2) is the monic Hermite polynomial of degree d, i.e. $\sqrt{d!}h_d(x) = x^d + \ldots$ where the leading term has coefficient 1, and that $\left(\frac{\sum_{i=1}^n z_i}{\sqrt{n}}\right)^d$ equals the right side of eq. (2) except without the restriction that the indices of summation $j_1, \ldots, j_d$ are distinct. The terms of the summation in which the indices $j_1, \ldots, j_d$ are not distinct are essentially cancelled out by the lower degree terms of $\sqrt{d!}h_d(x)$. Intuition for this cancellation comes from the observation that if $z_1, \ldots, z_n \overset{iid}{\sim} \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$, and if $\sqrt{d!}h_d$ is replaced by the monic degree-d orthogonal polynomial with respect to the distribution of $\frac{\sum_{i=1}^n z_i}{\sqrt{n}}$, then eq. (2) is actually an exact equality. This follows from the well-known fact that

$$p_d(z_1, \ldots, z_n) := \sum_{\substack{j_1,\ldots,j_d=1 \\ j_1\neq j_2\neq\ldots\neq j_d}}^{n} \prod_{i=1}^{d} z_{j_i}$$

satisfies $\mathbb{E}[p_d(z_1, \ldots, z_n)p_{d'}(z_1, \ldots, z_n)] = 0$ for any $d \neq d'$. When $z_1, \ldots, z_n$ are not Bernoulli-distributed and h_d is the Hermite polynomial, then eq. (2) is only an approximation, where the right side of eq. (2) may be considered as a first-order term of the left side. In Proposition 3.1, we quantify the error of this approximation by also computing

the second-order term, which is of size $O\left(\frac{1}{\sqrt{n}}\right)$ with high probability, and showing that the third and higher-order terms in this approximation are of size $O\left(\frac{1}{n}\right)$ with high probability.

(2) As $k(x) = \sum_{d=1}^{D} a_d h_d(x)$, the decomposition in step (1) above yields a corresponding decomposition $K_{n,p}(X) = Q_{n,p}(X) + R_{n,p}(X) + S_{n,p}(X)$ of the kernel random matrix, where Q, R, and S correspond to the first-order, second-order, and third-and-higher-order terms, respectively, of the decompositions in step (1) of the hermite polynomials $h_1, \dots, h_D$. The bulk of our argument lies in establishing that $\limsup_{n,p\to\infty} \|Q_{n,p}(X)\| \le \|\mu_{a,\nu,\gamma}\|$ almost surely. To prove this result, we use the moment method [15, 16]: For even integers l, $\|Q_{n,p}(X)\|^l \le \operatorname{Tr} Q_{n,p}(X)^l$, and a sufficiently tight upper bound may be obtained by taking $l := l(n) \asymp \log n$. Since $\mu_{a,\nu,\gamma}$ is the free additive convolution of a semicircle law with a scaled and translated Marcenko-Pastur law (Proposition 2.11), it is the limiting empirical spectral measure of a deformed GUE matrix of the form $M_{\tilde{n},\tilde{p}} = \sqrt{\frac{\gamma(\nu-a^2)}{\tilde{p}}} W_{\tilde{p}} + \frac{a}{\tilde{n}} V_{\tilde{n},\tilde{p}}$, as $\tilde{n}, \tilde{p} \to \infty$ with $\frac{\tilde{p}}{\tilde{n}} \to \gamma$, where $W_{\tilde{p}}$ is a $\tilde{p} \times \tilde{p}$ GUE matrix and $V_{\tilde{n},\tilde{p}}$ is an independent $\tilde{p} \times \tilde{p}$ sample covariance matrix based on $\tilde{n}$ samples and having zero diagonal. We employ combinatorial arguments to upper-bound the quantity $\mathbb{E}[\operatorname{Tr} Q_{n,p}(X)^l]$ using $\mathbb{E}[\operatorname{Tr} M_{\tilde{n},\tilde{p}}^l]$ for a suitable choice of $\tilde{n}$ and $\tilde{p}$, and we bound the latter quantity using the known convergence result $\lim_{\tilde{n},\tilde{p}\to\infty} \|M_{\tilde{n},\tilde{p}}\| = \|\mu_{a,\nu,\gamma}\|$.

(3) Finally, we analyze the remainder matrices $R_{n,p}(X)$ and $S_{n,p}(X)$ from the decomposition in step (2) above. It is easily shown that $\limsup_{n,p\to\infty} \|S_{n,p}(X)\| = 0$. For $R_{n,p}(X)$, we may write $R_{n,p}(X) = \sum_{d=2}^{D} R_{n,p,d}(X)$ where $R_{n,p,d}(X)$ is the contribution from the Hermite polynomial h_d of degree d. (The linear polynomial h_1 does not have such a remainder term in the decomposition from step (1).) Again using a moment bound, we show that $\limsup_{n,p\to\infty} \|R_{n,p,d}(X)\| = 0$ for each $d \ge 3$. For $d = 2$, we show that $\lim_{n,p\to\infty} \|R_{n,p,2}(X) - \tilde{R}_{n,p}(X)\| = 0$, where $\tilde{R}_{n,p}(X)$ is the rank-two matrix in Theorem 2.6. This establishes that $K_{n,p}(X) = \tilde{K}_{n,p}(X) + \tilde{R}_{n,p}(X)$, where $\limsup_{n,p\to\infty} \|\tilde{K}_{n,p}(X)\| \le \|\mu_{a,\nu,\gamma}\|$. The reverse inequality $\liminf_{n,p\to\infty} \|\tilde{K}_{n,p}(X)\| \ge \|\mu_{a,\nu,\gamma}\|$ follows immediately from Theorem 2.5, hence establishing Theorem 2.6.

Step (1) above is carried out in Section 3. Step (2) is carried out in Section 4, with many details of the combinatorial argument deferred to Appendix A and the estimate of $\mathbb{E}[\operatorname{Tr} M_{\tilde{n},\tilde{p}}^l]$ deferred to Appendix B. Step (3) and the conclusion of the proof of Theorem 2.6 are carried out in Section 5.

In Section 6, we prove Theorem 2.10 using Theorem 2.6. Our argument uses an approximation, in a suitable sense, of the kernel function $k(x)$ by a polynomial function $q(x)$. The spectral norm of the kernel matrix corresponding to $q(x)$ is easily estimated by Theorem 2.6, and we use a concentration of measure argument to control the spectral norm of the kernel matrix corresponding to the remainder $r(x) = k(x) - q(x)$. The concentration of measure argument relies on the construction of a certain dyadic covering net of the unit ball in $\mathbb{R}^p$ and is inspired by a similar argument of Rafal Latala [19].

3. Decomposition of Hermite polynomials of sums of IID random variables

The goal of this section is to prove the following proposition, which exhibits a three-term decomposition of a Hermite polynomial applied to a normalized sum $\frac{1}{\sqrt{n}} \sum_{j=1}^{n} z_j$ of i.i.d. random variables with zero mean, unit variance, and zero third moment.

Proposition 3.1. *Let $Z = (z_j : 1 \le j \le n) \in \mathbb{R}^n$, where the entries z_j are a collection of independent and identically distributed random variables such that $\mathbb{E}[z_j] = 0$, $\mathbb{E}[z_j^2] = 1$, $\mathbb{E}[z_j^3] = 0$, and $\mathbb{E}[|z_j|^l] < \infty$ for each $l \ge 1$. Let h_d denote the orthonormal Hermite polynomial of degree d as*

in Definition 2.4. Define

$$q_{d,n}(Z) = \sqrt{\frac{1}{n^d d!}} \sum_{\substack{j_1,\dots,j_d=1 \\ j_1 \neq j_2 \neq \dots \neq j_d}}^{n} \prod_{i=1}^{d} z_{j_i}, \tag{3}$$

$$r_{d,n}(Z) = \begin{cases} 0 & d = 1 \\ \sqrt{\frac{1}{n^d d!}} \binom{d}{2} \sum_{\substack{j_1,\dots,j_{d-1}=1 \\ j_1 \neq j_2 \neq \dots \neq j_{d-1}}}^{n} \left((z_{j_1}^2 - 1) \prod_{i=2}^{d-1} z_{j_i} \right) & d \geq 2 \,, \end{cases} \tag{4}$$

$$s_{d,n}(Z) = h_d\left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} z_j\right) - q_{d,n}(Z) - r_{d,n}(Z). \tag{5}$$

Then, for each $d \geq 1$ and any $\alpha, \beta > 0$, $\mathbb{P}[|s_{d,n}(Z)| > n^{-1+\alpha}] < n^{-\beta}$ for all sufficiently large n (i.e. for $n \geq N$ where N may depend on α, β, d, and the distribution of z_j).

Our proof of Proposition 3.1 proceeds via induction on d, using the three-term recurrence satisfied by the Hermite polynomials h_d, and it is presented at the end of this section. By Lemma 3.4 below, $\mathbb{P}\left[|q_{d,n}(Z)| > n^{\alpha}\right] < n^{-\beta}$ and $\mathbb{P}\left[|r_{d,n}(Z)| > n^{-\frac{1}{2}+\alpha}\right] < n^{-\beta}$ for any $\alpha, \beta > 0$ and all sufficiently large n. Hence Proposition 3.1 may be interpreted as decomposing $h_d\left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} z_j\right)$ as the sum of an "$O(1)$ term", an "$O\left(\frac{1}{\sqrt{n}}\right)$ term", and an "$O\left(\frac{1}{n}\right)$ term". Corresponding to this decomposition of the Hermite polynomials, let us consider the following decomposition of the kernel inner product matrix.

Definition 3.2. *Suppose $k(x) = \sum_{d=1}^{D} a_d h_d(x)$, as in Theorem 2.6. Then let $Q_{n,p}(X) = (q_{ii'} : 1 \leq i, i' \leq p) \in \mathbb{R}^{p \times p}$, with entries*

$$q_{ii'} = \begin{cases} \frac{1}{\sqrt{n}} \sum_{d=1}^{D} a_d q_{d,n}(x_{i1}x_{i'1}, \dots, x_{in}x_{i'n}), & i \neq i' \\ 0, & i = i', \end{cases}$$

where $q_{d,n}$ is defined as in eq. (3). Let $R_{n,p} \in \mathbb{R}^{p \times p}$ and $S_{n,p} \in \mathbb{R}^{p \times p}$ be defined analogously using the functions $r_{d,n}$ and $s_{d,n}$, from eqs. (4) and (5), respectively, in place of $q_{d,n}$.

Remark 3.3. *By Definitions 2.2 and 3.2 and the definition of $s_{d,n}$ in eq. (5), it is evident that*

$$K_{n,p}(X) = Q_{n,p}(X) + R_{n,p}(X) + S_{n,p}(X),$$

where $K_{n,p}(X)$ is as in Theorem 2.6.

Lemma 3.4. *Suppose $z_1, \dots, z_n$ are i.i.d. random variables, with $\mathbb{E}[|z_j|^l] < \infty$ for all $l \geq 1$. Let $p_1, \dots, p_d : \mathbb{R} \to \mathbb{R}$ be any polynomial functions such that $\mathbb{E}[p_i(z_j)] = 0$ for each $i = 1, \dots, d$. Then for any $\alpha, \beta > 0$,*

$$\mathbb{P}\left[n^{-\frac{d}{2}} \left| \sum_{\substack{j_1,\dots,j_d=1 \\ j_1 \neq j_2 \neq \dots \neq j_d}}^{n} \prod_{i=1}^{d} p_i(z_{j_i}) \right| > n^{\alpha} \right] < n^{-\beta}$$

for all sufficiently large n (i.e. $n \geq N$ where N may depend on α, β, d and the distribution of z_j).

Proof. Fix $\alpha, \beta > 0$. Let

$$f(z_1, \dots, z_n) = n^{-\frac{d}{2}} \left| \sum_{\substack{j_1,\dots,j_d=1 \\ j_1 \neq j_2 \neq \dots \neq j_d}}^{n} \prod_{i=1}^{d} p_i(z_{j_i}) \right|,$$

and let l be an even integer such that $\alpha l > \beta$. Then

$$\mathbb{P}[f(z_1, \dots, z_n) > n^{\alpha}] \leq \frac{\mathbb{E}[f(z_1, \dots, z_n)^l]}{n^{\alpha l}},$$

and it suffices to show $\mathbb{E}[f(z_1, \dots, z_n)^l] \leq C$ for a constant C independent of n. Note that

$$\mathbb{E}[f(z_1, \dots, z_n)^l] = n^{-\frac{ld}{2}} \sum_{\substack{j_1^1,\dots,j_d^1=1 \\ j_1^1 \neq \dots \neq j_d^1}}^{n} \dots \sum_{\substack{j_1^l,\dots,j_d^l=1 \\ j_1^l \neq \dots \neq j_d^l}}^{n} \mathbb{E}\left[\prod_{i=1}^{d} \prod_{k=1}^{l} p_i\left(z_{j_i^k}\right) \right].$$

For each term of the above sum, if there is some j such that $j = j_i^k$ for exactly one pair of indices $i \in \{1, \dots, d\}$ and $k \in \{1, \dots, l\}$, then the expectation of that term is 0 as $\mathbb{E}[p_i(z_j)] = 0$ and z_j is independent of $z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_n$. Hence, for terms in the sum with non-zero expectation, there are at most $\frac{ld}{2}$ distinct values of j_i^k. Then the number of such terms is at most $C_{l,d}\binom{n}{\frac{ld}{2}}$ for a combinatorial constant $C_{l,d}$ not depending on n. Also, the magnitude of the expectation of each such term is at most $C' < \infty$ for a constant C' depending on l, d, the polynomials $p_1, \dots, p_d$, and the absolute moments of z_j (which are finite by assumption). As $\binom{n}{\frac{ld}{2}} \leq n^{\frac{ld}{2}}$, the result follows. □

Proof of Proposition 3.1. Let $S = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} z_j$. It will be notationally convenient to work with the monic Hermite polynomials $\tilde{h}_d = h_d \sqrt{d!}$, so that $\tilde{h}_d$ has leading coefficient 1. Let us accordingly define $\tilde{q}_{d,n} = q_{d,n}\sqrt{d!}$, $\tilde{r}_{d,n} = r_{d,n}\sqrt{d!}$, and $\tilde{s}_{d,n} = s_{d,n}\sqrt{d!}$. Then

$$\tilde{h}_d(S) = \tilde{q}_{d,n}(Z) + \tilde{r}_{d,n}(Z) + \tilde{s}_{d,n}(Z),$$

and we wish to show for any $\alpha, \beta > 0$, $\mathbb{P}\left[|\tilde{s}_d(Z)| > n^{-1+\alpha}\right] < n^{-\beta}$ for all sufficiently large n.

We proceed by induction on d. Note that $\tilde{h}_0(x) = 1$, $\tilde{h}_1(x) = x$, and $\tilde{h}_2(x) = x^2 - 1$. Then for $d = 1$, $\tilde{h}_1(S) = S = \tilde{q}_{1,n}(Z)$, and for $d = 2$,

$$\tilde{h}_2(S) = S^2 - 1 = n^{-1} \left(\sum_{\substack{j_1,j_2=1 \\ j_1 \neq j_2}}^{n} z_{j_1} z_{j_2} + \sum_{j=1}^{n} (z_j^2 - 1) \right) = \tilde{q}_{2,n}(Z) + \tilde{r}_{2,n}(Z).$$

Hence the proposition holds with $\tilde{s}_{1,n}(Z) = \tilde{s}_{2,n}(Z) = 0$.

Let us assume by induction that the proposition holds for $d-1$ and d. Recall that the monic Hermite polynomials satisfy the three-term recurrence $\tilde{h}_{d+1}(x) = x\tilde{h}_d(x) - d\tilde{h}_{d-1}(x)$ (c.f. eq. (5.5.8) of [28]). We may compute

$$\begin{aligned} S\tilde{q}_{d,n}(Z) &= n^{-\frac{d+1}{2}} \sum_{j=1}^{n} z_j \sum_{\substack{j_1,\dots,j_d=1 \\ j_1 \neq \dots \neq j_d}}^{n} \prod_{i=1}^{d} z_{j_i} \\ &= n^{-\frac{d+1}{2}} \left(\sum_{\substack{j_1,\dots,j_{d+1}=1 \\ j_1 \neq \dots \neq j_{d+1}}}^{n} \prod_{i=1}^{d+1} z_{j_i} + d \sum_{\substack{j_1,\dots,j_d=1 \\ j_1 \neq \dots \neq j_d}}^{n} z_{j_1}^2 \prod_{i=2}^{d} z_{j_i} \right) \end{aligned}$$

$$
\begin{aligned}
&= \tilde{q}_{d+1,n}(Z) + \frac{2}{d+1}\tilde{r}_{d+1,n}(Z) + \frac{d(n-d+1)}{n}\tilde{q}_{d-1,n}(Z),\\
S\tilde{r}_{d,n}(Z) &= n^{-\frac{d+1}{2}}\binom{d}{2}\sum_{j=1}^{n} z_j \sum_{\substack{j_1,\dots,j_{d-1}=1\\ j_1\neq\dots\neq j_{d-1}}}^{n}\left((z_{j_1}^2-1)\prod_{i=2}^{d-1} z_{j_i}\right)\\
&= n^{-\frac{d+1}{2}}\binom{d}{2}\left(\sum_{\substack{j_1,\dots,j_d=1\\ j_1\neq\dots\neq j_d}}^{n}\left((z_{j_1}^2-1)\prod_{i=2}^{d} z_{j_i}\right) + \sum_{\substack{j_1,\dots,j_{d-1}=1\\ j_1\neq\dots\neq j_{d-1}}}^{n}\left((z_{j_1}^3-z_{j_1})\prod_{i=2}^{d-1} z_{j_i}\right)\right.\\
&\qquad \left. +(d-2)\sum_{\substack{j_1,\dots,j_{d-1}=1\\ j_1\neq\dots\neq j_{d-1}}}^{n}\left((z_{j_1}^2-1)z_{j_2}^2\prod_{i=3}^{d-1} z_{j_i}\right)\right)\\
&= \frac{d-1}{d+1}\tilde{r}_{d+1,n}(Z) + n^{-\frac{d+1}{2}}\binom{d}{2}\sum_{\substack{j_1,\dots,j_{d-1}=1\\ j_1\neq\dots\neq j_{d-1}}}^{n}\left((z_{j_1}^3-z_{j_1})\prod_{i=2}^{d-1} z_{j_i}\right)\\
&\qquad + n^{-\frac{d+1}{2}}\binom{d}{2}(d-2)\sum_{\substack{j_1,\dots,j_{d-1}=1\\ j_1\neq\dots\neq j_{d-1}}}^{n}(z_{j_1}^2-1)(z_{j_2}^2-1)\prod_{i=3}^{d-1} z_{j_i} + \frac{d(n-d+2)}{n}\tilde{r}_{d-1,n}(Z).
\end{aligned}
$$

Substituting these expressions into the three-term recurrence,

$$
\begin{aligned}
\tilde{h}_{d+1}(S) &= S\left(\tilde{q}_{d,n}(Z)+\tilde{r}_{d,n}(Z)+\tilde{s}_{d,n}(Z)\right) - d\left(\tilde{q}_{d-1}(Z)+\tilde{r}_{d-1}(Z)+\tilde{s}_{d-1}(Z)\right)\\
&= \tilde{q}_{d+1,n}(Z)+\tilde{r}_{d+1,n}(Z)+\tilde{s}_{d+1,n}(Z)
\end{aligned}
$$

for

$$
\begin{aligned}
\tilde{s}_{d+1,n}(Z) &:= -\frac{d(d-1)}{n}\tilde{q}_{d-1,n}(Z) + n^{-\frac{d+1}{2}}\binom{d}{2}\sum_{\substack{j_1,\dots,j_{d-1}=1\\ j_1\neq\dots\neq j_{d-1}}}^{n}\left((z_{j_1}^3-z_{j_1})\prod_{i=2}^{d-1} z_{j_i}\right)\\
&\quad + n^{-\frac{d+1}{2}}\binom{d}{2}(d-2)\sum_{\substack{j_1,\dots,j_{d-1}=1\\ j_1\neq\dots\neq j_{d-1}}}^{n}\left((z_{j_1}^2-1)(z_{j_2}^2-1)\prod_{i=3}^{d-1} z_{j_i}\right) - \frac{d(d-2)}{n}\tilde{r}_{d-1,n}(Z)\\
&\quad + S\tilde{s}_{d,n}(Z) - d\tilde{s}_{d-1,n}(Z)\\
&=: I + II + III + IV + V + VI.
\end{aligned}
$$

Fix $\alpha,\beta>0$. Note that $\mathbb{E}[z_j]=0$, $\mathbb{E}[z_j^2-1]=0$, and $\mathbb{E}[z_j^3-z_j]=0$, so by Lemma 3.4,

$$
\max\left(\mathbb{P}\left[|I|>n^{-1+\frac{\alpha}{2}}\right],\mathbb{P}\left[|II|>n^{-1+\frac{\alpha}{2}}\right],\mathbb{P}\left[|III|>n^{-1+\frac{\alpha}{2}}\right],\mathbb{P}\left[|IV|>n^{-1+\frac{\alpha}{2}}\right]\right) < n^{-2\beta}
$$

for all large n. By the induction hypothesis, $\mathbb{P}\left[|\tilde{s}_{d,n}|>n^{-1+\frac{\alpha}{4}}\right] < \frac{n^{-2\beta}}{2}$ for all large n, and also $\mathbb{P}\left[|S|>n^{\frac{\alpha}{4}}\right] < \frac{n^{-2\beta}}{2}$ for all large n by Lemma 3.4 (applied to the simple case where $d=1$ and $p_1(x)=x$). Then $\mathbb{P}\left[|V|>n^{-1+\frac{\alpha}{2}}\right] < n^{-2\beta}$ for all large n. Similarly, the induction hypothesis

implies $\mathbb{P}\left[|VI| > n^{-1+\frac{\alpha}{2}}\right] < n^{-2\beta}$ for all large n. Putting this together,

$$\begin{aligned}\mathbb{P}\left[|\tilde{s}_{d+1,n}(Z)| > n^{-1+\alpha}\right] &\leq \mathbb{P}\left[|I + II + III + IV + V + VI| > 6n^{-1+\frac{\alpha}{2}}\right] \\ &\leq \mathbb{P}\left[|I| > n^{-1+\frac{\alpha}{2}}\right] + \ldots + \mathbb{P}\left[|VI| > n^{-1+\frac{\alpha}{2}}\right] \\ &< 6n^{-2\beta} < n^{-\beta}\end{aligned}$$

for all large n, completing the induction. □

4. Bounding the dominant term $\|Q_{n,p}(X)\|$

In this section, we establish the following result.

Proposition 4.1. *Let $Q_{n,p}(X)$ be as in Definition 3.2. Let $\mu_{a,\nu,\gamma}$ be as in Definition 2.3, with a, ν, γ as specified in Theorem 2.6. Then $\limsup_{n,p\to\infty} \|Q_{n,p}(X)\| \leq \|\mu_{a,\nu,\gamma}\|$ almost surely.*

Our proof uses the moment method. The following definition of a multi-labeling of an l-graph will correspond to the primary combinatorial object of interest in the subsequent analysis.

Definition 4.2. *For any integer $l \geq 2$, an **l-graph** is a graph consisting of a single cycle with $2l$ vertices and $2l$ edges, with the vertices alternatingly denoted as **p-vertices** and **n-vertices**.*

We will consider the vertices of the l-graph to be ordered by picking an arbitrary p-vertex as the first vertex and ordering the remaining vertices according to a traversal along the cycle. A vertex V "follows" or "precedes" another vertex W if V comes before or after W, respectively, in this ordering, and the last vertex of the cycle (which is an n-vertex) is followed by the first p-vertex.

Definition 4.3. *A **multi-labeling** of an l-graph is an assignment of a **p-label** in $\{1, 2, 3, \ldots\}$ to each p-vertex and an ordered tuple of **n-labels** in $\{1, 2, 3, \ldots\}$ to each n-vertex, such that the following conditions are satisfied:*

1. *The p-label of each p-vertex is distinct from those of the two p-vertices immediately preceding and following it in the cycle.*
2. *The number d_s of n-labels in the tuple for each s^{th} n-vertex satisfies $1 \leq d_s \leq D$, and these d_s n-labels are distinct.*
3. *For each distinct p-label i and distinct n-label j, there are an even number of edges in the cycle (possibly 0) such that its p-vertex endpoint is labeled i and its n-vertex endpoint has label j in its tuple.*

*A **(p, n)-multi-labeling** is a multi-labeling with all p-labels in $\{1, \ldots, p\}$ and all n-labels in $\{1, \ldots, n\}$.*

Figure 2 shows an example of a $(3, 5)$-multi-labeling of an l-graph, for $l = 4$ and $D = 3$. In Definition 4.3 and the subsequent combinatorial arguments, D corresponds to the degree of the polynomial k in Theorem 2.6, which we will always treat as a fixed quantity. We will always consider p-labels to be distinct from n-labels, even though (for notational convenience) we use the same label set $\{1, 2, 3, \ldots\}$ for both.

A key bound on the number of possible distinct p-labels and n-labels that appear in a multi-labeling of an l-graph is provided by the following lemma.

Lemma 4.4. *Suppose a multi-labeling of an l-graph has $d_1, \ldots, d_l$ n-labels on the first through l^{th} n-vertices, respectively, and suppose that it has m total distinct p-labels and n-labels. Then $m \leq \frac{l+\sum_{s=1}^{l} d_s}{2} + 1$.*

We defer the proof of Lemma 4.4 to Appendix A. The quantity $\frac{l+\sum_{s=1}^{l} d_s}{2} + 1 - m$ will appear in many of our subsequent combinatorial lemmas, and we give it a name.

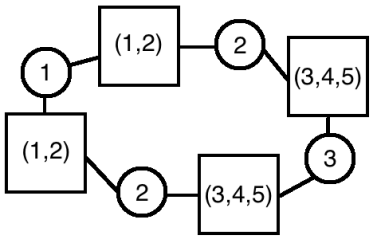

FIGURE 2. A $(3,5)$-multi-labeling of an l-graph, for $l = 4$ and $D = 3$. p-vertices are depicted with a circle and n-vertices are depicted with a square.

Definition 4.5. *Suppose a multi-labeling of an l-graph has $d_1, \ldots, d_l$ n-labels on the first through l^{th} n-vertices, respectively, and suppose that it has m total distinct p-labels and n-labels. The* **excess** *of the multi-labeling is $\Delta := \frac{l+\sum_{s=1}^{l} d_s}{2} + 1 - m$.*

Definition 4.6. *Two multi-labelings of an l-graph are* **equivalent** *if there is a permutation π_p of $\{1,2,3,\ldots\}$ and a permutation π_n of $\{1,2,3,\ldots\}$ such that one labeling is the image of the other upon applying π_p to all of its p-labels and π_n to all of its n-labels. For any fixed l, the equivalence classes under this relation will be called* **multi-labeling equivalence classes**.

Note that Lemma 4.4 implies that the excess Δ of a multi-labeling is always nonnegative. Under these definitions, the number of distinct p-labels, number of distinct n-labels, number of n-labels $d_1, \ldots, d_l$ for each of the l n-vertices, and excess Δ are equivalence class properties, i.e. they are the same for all labelings in the same multi-labeling equivalence class. The motivation for Definition 4.3 of a multi-labeling is provided by the following lemma.

Lemma 4.7. *Let $Q_{n,p}(X)$ be as in Proposition 4.1, and let $l \geq 2$ be an even integer. Let $\mathcal{C}$ denote the set of all multi-labeling equivalence classes for an l-graph. For each multi-labeling equivalence class $\mathcal{L} \in \mathcal{C}$, let $\Delta(\mathcal{L})$ be the excess, $r(\mathcal{L})$ the number of distinct p-labels, and $d_1(\mathcal{L}), \ldots, d_l(\mathcal{L})$ the number of n-labels on the first to l^{th} n-vertices, respectively. Then, with $\alpha > 0$ as in Assumption 2.1 and with the convention $0^0 = 1$,*

$$\mathbb{E}[\operatorname{Tr} Q_{n,p}(X)^l] \leq n \sum_{\mathcal{L} \in \mathcal{C}} \left(\frac{(12\Delta(\mathcal{L}))^{12\alpha}}{n} \right)^{\Delta(\mathcal{L})} \left(\frac{p}{n} \right)^{r(\mathcal{L})} \left(\prod_{s=1}^{l} \frac{a_{d_s(\mathcal{L})}}{(d_s(\mathcal{L})!)^{1/2}} \right). \tag{6}$$

Proof. By Definition 3.2, letting $i_{l+1} := i_1$ for notational convenience,

$$\begin{aligned}
\mathbb{E}[\operatorname{Tr} Q_{n,p}(X)^l] &= \sum_{\substack{i_1,\ldots,i_l=1 \\ i_1 \neq i_2, i_2 \neq i_3, \ldots, i_l \neq i_1}}^{p} \mathbb{E}\left[\prod_{s=1}^{l} q_{i_s i_{s+1}} \right] \\
&= \sum_{\substack{i_1,\ldots,i_l=1 \\ i_1 \neq i_2, i_2 \neq i_3, \ldots, i_l \neq i_1}}^{p} n^{-\frac{l}{2}} \mathbb{E}\left[\prod_{s=1}^{l} \left(\sum_{d=1}^{D} a_d \sqrt{\frac{1}{n^d d!}} \sum_{\substack{j_1,\ldots,j_d=1 \\ j_1 \neq j_2 \neq \ldots \neq j_d}}^{n} \prod_{a=1}^{d} x_{i_s j_a} x_{i_{s+1} j_a} \right) \right] \\
&= \sum_{\substack{i_1,\ldots,i_l=1 \\ i_1 \neq i_2, i_2 \neq i_3, \ldots, i_l \neq i_1}}^{p} \sum_{d_1,\ldots,d_s=1}^{D} \sum_{\substack{j_1^1,\ldots,j_{d_1}^1=1 \\ j_1^1 \neq \ldots \neq j_{d_1}^1}}^{n} \cdots \sum_{\substack{j_1^l,\ldots,j_{d_l}^l=1 \\ j_1^l \neq \ldots \neq j_{d_l}^l}}^{n} \\
&\qquad n^{-\frac{l+\sum_{s=1}^{l} d_s}{2}} \left(\prod_{s=1}^{l} \frac{a_{d_s}}{(d_s!)^{1/2}} \right) \mathbb{E}\left[\prod_{s=1}^{l} \prod_{a=1}^{d_s} x_{i_s j_a^s} x_{i_{s+1} j_a^s} \right].
\end{aligned}$$

Note that as $x_{ij} \stackrel{L}{=} -x_{ij}$ by Assumption 2.1, $\mathbb{E}[x_{ij}^c] = 0$ for any positive odd integer c. Hence, if any x_{ij} appears an odd number of times in the expression $\prod_{s=1}^{l}\prod_{a=1}^{d_s} x_{i_s j_a^s} x_{i_{s+1} j_a^s}$, then, as x_{ij} is independent of $x_{i'j'}$ if $j \neq j'$ or $i \neq i'$, $\mathbb{E}\left[\prod_{s=1}^{l}\prod_{a=1}^{d_s} x_{i_s j_a^s} x_{i_{s+1} j_a^s}\right] = 0$ for such a term. We identify the combination of the sums above, over the remaining non-zero terms, as the sum over all possible (p,n)-multi-labelings of an l-graph. Here, the first sum over $i_1, \ldots, i_l$ is over all choices of p-labels, with condition (1) in Definition 4.3 that any two consecutive p-vertices have distinct labels being imposed by the constraints $i_1 \neq i_2, i_2 \neq i_3, \ldots, i_l \neq i_1$ in the sum. The sum over $d_1, \ldots, d_s$ is over all possible choices of the number of n-labels in the n-label tuple for each n-vertex, and the sum over $j_1^s, \ldots, j_{d_s}^s$ is over all possible choices of d_s n-labels for the s^{th} n-vertex, with condition (2) in Definition 4.3 that the p-labels for each p-vertex are distinct being imposed by the constraint that $j_1^s, \ldots, j_{d_s}^s$ are distinct. The product expression $\prod_{s=1}^{l}\prod_{a=1}^{d_s} x_{i_s j_a^s} x_{i_{s+1} j_a^s}$ then corresponds to a product, over all n-vertices, all d_s n-labels for that n-vertex, and both p-vertices immediately preceding and immediately following that n-vertex, of x_{ij}, where $j \in \{1, \ldots, n\}$ is the n-label and $i \in \{1, \ldots, p\}$ is the p-label of the p-vertex. The condition that x_{ij} for each distinct pair (i,j) appears an even number of times, so that this term has non-zero expectation, is precisely condition (3) in Definition 4.3. Thus, to summarize,

$$\mathbb{E}[\operatorname{Tr} Q_{n,p}(X)^l] = \sum_{l\text{-graph } (p,n)\text{-multi-labelings}} n^{-\frac{l+\sum_{s=1}^{l} d_s}{2}} \left(\prod_{s=1}^{l} \frac{a_{d_s}}{(d_s!)^{1/2}}\right) \mathbb{E}\left[\prod_{s=1}^{l}\prod_{a=1}^{d_s} x_{i_s j_a^s} x_{i_{s+1} j_a^s}\right],$$

where $d_1, \ldots, d_l$ are the numbers of n-labels for the first through l^{th} n-vertices, respectively.

Consider a fixed (p,n)-multi-labeling and write $\prod_{s=1}^{l}\prod_{a=1}^{d_s} x_{i_s j_a^s} x_{i_{s+1} j_a^s} = \prod_{j=1}^{n}\prod_{i=1}^{p} x_{ij}^{b_{ij}}$, where b_{ij} is the number of times x_{ij} appears as a term in this product. Note that each b_{ij} is even (possibly 0). As $\mathbb{E}[x_{ij}^2] = 1$, $\mathbb{E}[|x_{ij}|^k] \leq k^{\alpha k}$, and x_{ij} is independent of $x_{i'j'}$ if $j' \neq j$ or $i' \neq i$,

$$\mathbb{E}\left[\prod_{s=1}^{l}\prod_{a=1}^{d_s} x_{i_s j_a^s} x_{i_{s+1} j_a^s}\right] = \prod_{i,j:b_{ij}>2} \mathbb{E}\left[x_{ij}^{b_{ij}}\right] \leq \prod_{i,j:b_{ij}>2} b_{ij}^{\alpha b_{ij}} \leq \left(\sum_{i,j:b_{ij}>2} b_{ij}\right)^{\alpha \sum_{i,j:b_{ij}>2} b_{ij}},$$

where the last inequality holds with the convention $0^0 = 1$ if $b_{ij} \leq 2$ for all (i,j). We show in Lemma A.6 that $\sum_{i,j:b_{ij}>2} b_{ij} \leq 12\Delta$, so $\mathbb{E}\left[\prod_{s=1}^{l}\prod_{a=1}^{d_s} x_{i_s j_a^s} x_{i_{s+1} j_a^s}\right] \leq (12\Delta)^{12\alpha\Delta}$. Eq. (6) then follows upon noting that each (p,n)-multi-labeling with r distinct p-labels and $m-r$ distinct n-labels has $\frac{p!}{(p-r)!}\frac{n!}{(n-m+r)!} \leq n^m \left(\frac{p}{n}\right)^r$ (p,n)-multi-labelings in its equivalence class, and $n^{-\frac{l+\sum_{s=1}^{l} d_s}{2}+m} = n^{1-\Delta}$. □

To prove Proposition 4.1, it remains to control the upper bound in eq. (6). We do so by comparing this quantity to an analogous quantity for a deformed GUE matrix, specified in the following definition.

Definition 4.8. *For $\tilde{n}, \tilde{p} \geq 1$, $\nu, \gamma > 0$, and $a \in [-\sqrt{\nu}, \sqrt{\nu}]$, let $W_{\tilde{p}} = (w_{ii'} : 1 \leq i, i' \leq \tilde{p}) \in \mathbb{C}^{\tilde{p}\times\tilde{p}}$ be distributed according to the GUE, i.e. $\{w_{ii} : 1 \leq i \leq \tilde{p}\} \cup \{\sqrt{2}\operatorname{Re} w_{ii'}, \sqrt{2}\operatorname{Im} w_{ii'} : 1 \leq i < i' \leq \tilde{p}\}$ are i.i.d. $\mathcal{N}(0,1)$, and $w_{ii'} = \overline{w_{i'i}}$ for $i > i'$. Let $V_{\tilde{n},\tilde{p}} \in \mathbb{R}^{\tilde{p}\times\tilde{p}}$ be standard real Wishart-distributed with $\tilde{n}$ degrees of freedom and zero diagonal, i.e. $V_{\tilde{n},\tilde{p}} = ZZ^T - \operatorname{diag}(\|Z_i\|_2^2)$ where $Z := Z_{\tilde{n},\tilde{p}} = (z_{ij} : 1 \leq i \leq \tilde{p}, 1 \leq j \leq \tilde{n}) \in \mathbb{R}^{\tilde{p}\times\tilde{n}}$, $z_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$, and $\operatorname{diag}(\|Z_i\|_2^2) \in \mathbb{R}^{\tilde{p}\times\tilde{p}}$ denotes the diagonal matrix whose i^{th} diagonal entry is the squared Euclidean norm of the i^{th} row of Z. Take $V_{\tilde{n},\tilde{p}}$ and $W_{\tilde{p}}$ to be independent, and let $M_{\tilde{n},\tilde{p}} = \sqrt{\frac{\gamma(\nu-a^2)}{\tilde{p}}} W_{\tilde{p}} + \frac{a}{\tilde{n}} V_{\tilde{n},\tilde{p}} \in \mathbb{C}^{\tilde{p}\times\tilde{p}}$.*

As $\tilde{n}, \tilde{p} \to \infty$ with $\frac{\tilde{p}}{\tilde{n}} \to \gamma$, the limiting spectral distribution of $M_{\tilde{n},\tilde{p}}$ is given by the free additive convolution of a scaled semicircle law and a scaled and translated Marcenko-Pastur law. We have

chosen the scaling factors for $W_{\tilde{p}}$ and $V_{\tilde{n},\tilde{p}}$ so that this free convolution is exactly the measure $\mu_{a,\nu,\gamma}$ in Definition 2.3. It follows from the results of [6] that, in fact, a norm convergence result holds for $M_{\tilde{n},\tilde{p}}$, i.e. $\lim_{\tilde{n},\tilde{p}\to\infty} \|M_{\tilde{n},\tilde{p}}\| = \|\mu_{a,\nu,\gamma}\|$, from which we may deduce the following Proposition.

Proposition 4.9. *Let $M_{\tilde{n},\tilde{p}}$ be as in Definition 4.8. Suppose l is an even integer and $\tilde{n}, \tilde{p}, l \to \infty$ with $\frac{\tilde{p}}{\tilde{n}} \to \gamma$ and $l \leq C \log \tilde{n}$ for some constant $C > 0$. Then, for any $\varepsilon > 0$,*

$$\mathbb{E}[\|M_{\tilde{n},\tilde{p}}\|^l] \leq (\|\mu_{a,\nu,\gamma}\| + \varepsilon)^l$$

for all sufficiently large $\tilde{n}$.

The proof of Proposition 4.9 is deferred to Appendix B. (Let us remark that our combinatorial argument would be somewhat simplified if, in Definition 4.8, we could replace the GUE matrix by a GOE matrix, but we cannot find a rigorous statement of the result $\lim_{n,p\to\infty} \|M_{\tilde{n},\tilde{p}}\| = \|\mu_{a,\nu,\gamma}\|$ a.s. for the GOE case in the literature.) As $\frac{1}{\tilde{p}}\mathbb{E}[\operatorname{Tr} M^l_{\tilde{n},\tilde{p}}] \leq \mathbb{E}[\|M_{\tilde{n},\tilde{p}}\|^l]$, our strategy for proving Proposition 4.1 will be to show that the upper bound in eq. (6) can in turn be bounded above using the quantity $\mathbb{E}[\operatorname{Tr} M^l_{\tilde{n},\tilde{p}}]$, for some choices of $\tilde{p}$ and $\tilde{n}$. To analyze the quantity $\mathbb{E}[\operatorname{Tr} M^l_{\tilde{n},\tilde{p}}]$, we consider the following notion of a simple-labeling of an l-graph.

Definition 4.10. *A* **simple-labeling** *of an l-graph is an assignment of a* ***p*-label** *in $\{1, 2, 3, \ldots\}$ to each p-vertex and either one* ***n*-label** *in $\{1, 2, 3, \ldots\}$ or the empty label $\emptyset$ to each n-vertex, such that the following conditions are satisfied:*

1. *The p-label of each p-vertex is distinct from those of the two p-vertices immediately preceding and following it in the cycle.*
2. *For each distinct p-label i and distinct non-empty n-label j, there are an even number of edges in the cycle (possibly 0) such that its p-vertex endpoint is labeled i and its n-vertex endpoint is labeled j.*
3. *For any two distinct p-labels i and i', the number of occurrences (possibly 0) of the three consecutive labels $i, \emptyset, i'$ on a p-vertex, its following n-vertex, and its following p-vertex, respectively, is equal to the number of occurrences of the three consecutive labels $i', \emptyset, i$.*

A **(p, n)-simple-labeling** *is a simple-labeling with all p-labels in $\{1, \ldots, p\}$ and all non-empty n-labels in $\{1, \ldots, n\}$.*

Analogous to Lemma 4.4, the following lemma provides a key bound on the number of possible distinct p-labels and n-labels that appear in a simple-labeling of an l-graph.

Lemma 4.11. *Suppose a simple-labeling of an l-graph has $\tilde{k}$ n-vertices with non-empty label and $\tilde{m}$ total distinct p-labels and distinct non-empty n-labels. Then $\tilde{m} \leq \frac{l+\tilde{k}}{2} + 1$.*

The proof of Lemma 4.11 is deferred to Appendix A. We may then define the excess of a simple-labeling, analogous to Definition 4.5, and note that the excess is always nonnegative.

Definition 4.12. *Suppose a simple-labeling of an l-graph has $\tilde{k}$ n-vertices with non-empty label and $\tilde{m}$ total distinct p-labels and distinct non-empty n-labels. The* **excess** *of the simple-labeling is $\tilde{\Delta} := \frac{l+\tilde{k}}{2} + 1 - \tilde{m}$.*

Definition 4.13. *Two simple-labelings of an l-graph are* **equivalent** *if there is a permutation π_p of $\{1, 2, 3, \ldots\}$ and a permutation π_n of $\{1, 2, 3, \ldots\}$ such that one labeling is the image of the other upon applying π_p to all of its p-labels and π_n to all of its n-labels. (The empty n-label remains empty under any such permutation π_n.) For any fixed l, the equivalence classes under this relation will be called* **simple-labeling equivalence classes**.

Motivation for Definition 4.10 of a simple labeling is provided by the following lemma, which gives a lower bound for the quantity $\mathbb{E}[\operatorname{Tr} M^l_{\tilde{n},\tilde{p}}]$ analogous to the upper bound for $\mathbb{E}[\operatorname{Tr} Q_{n,p}(X)^l]$ in Lemma 4.7.

Lemma 4.14. *Let $M_{\tilde{n},\tilde{p}}$ be as in Definition 4.8, and let $l \geq 2$ be an even integer. Let $\tilde{\mathcal{C}}$ denote the set of all simple-labeling equivalence classes for an l-graph. For each simple-labeling equivalence class $\tilde{\mathcal{L}} \in \tilde{\mathcal{C}}$, let $\tilde{\Delta}(\tilde{\mathcal{L}})$ be its excess, $\tilde{k}(\tilde{\mathcal{L}})$ be the number of n-vertices with non-empty label, and $\tilde{r}(\tilde{\mathcal{L}})$ be the number of distinct p-labels. Then, with the convention $0^0 = 1$,*

$$\mathbb{E}[\operatorname{Tr} M_{\tilde{n},\tilde{p}}^l] \geq \tilde{n} \left(\frac{\tilde{p}-l}{\tilde{p}}\right)^l \left(\frac{\tilde{n}-l}{\tilde{n}}\right)^l \sum_{\tilde{\mathcal{L}} \in \tilde{\mathcal{C}}} \left(\frac{1}{\tilde{n}}\right)^{\tilde{\Delta}(\tilde{\mathcal{L}})} \left(\frac{\tilde{p}}{\tilde{n}}\right)^{\tilde{r}(\tilde{\mathcal{L}}) - \frac{l-\tilde{k}(\tilde{\mathcal{L}})}{2}} |a|^{\tilde{k}(\tilde{\mathcal{L}})} (\gamma(\nu - a^2))^{\frac{l-\tilde{k}(\tilde{\mathcal{L}})}{2}}. \tag{7}$$

Proof. By Definition 4.8, letting $i_{l+1} := i_1$ for notational convenience,

$$\begin{aligned}
\mathbb{E}\left[\operatorname{Tr} M_{\tilde{n},\tilde{p}}^l\right] &= \mathbb{E}\left[\operatorname{Tr}\left(\sqrt{\frac{\gamma(\nu-a^2)}{\tilde{p}}} W_{\tilde{p}} + \frac{a}{\tilde{n}} V_{\tilde{n},\tilde{p}}\right)^l\right] \\
&= \sum_{i_1,\ldots,i_l=1}^{\tilde{p}} \mathbb{E}\left[\prod_{s=1}^{l}\left(\sqrt{\frac{\gamma(\nu-a^2)}{\tilde{p}}} w_{i_s i_{s+1}} + \frac{a}{\tilde{n}} v_{i_s i_{s+1}}\right)\right] \\
&= \sum_{i_1,\ldots,i_l=1}^{\tilde{p}} \sum_{S \subseteq \{1,\ldots,l\}} \left(\frac{a}{\tilde{n}}\right)^{|S|} \left(\frac{\gamma(\nu-a^2)}{\tilde{p}}\right)^{\frac{l-|S|}{2}} \mathbb{E}\left[\prod_{s\in S} v_{i_s i_{s+1}}\right] \mathbb{E}\left[\prod_{s\notin S} w_{i_s i_{s+1}}\right] \\
&= \sum_{S \subseteq \{1,\ldots,l\}} \sum_{\substack{i_1,\ldots,i_l=1 \\ i_s \neq i_{s+1} \forall s \in S}}^{\tilde{p}} \tilde{n}^{-\frac{l+|S|}{2}} \left(\frac{\tilde{p}}{\tilde{n}}\right)^{-\frac{l-|S|}{2}} a^{|S|} (\gamma(\nu-a^2))^{\frac{l-|S|}{2}} \mathbb{E}\left[\prod_{s\in S} v_{i_s i_{s+1}}\right] \mathbb{E}\left[\prod_{s\notin S} w_{i_s i_{s+1}}\right] \\
&= \sum_{S \subseteq \{1,\ldots,l\}} \sum_{\substack{i_1,\ldots,i_l=1 \\ i_s \neq i_{s+1} \forall s \in S}}^{\tilde{p}} \sum_{(j_s : s\in S) \in \{1,\ldots,\tilde{n}\}^{|S|}} \\
&\qquad \tilde{n}^{-\frac{l+|S|}{2}} \left(\frac{\tilde{p}}{\tilde{n}}\right)^{-\frac{l-|S|}{2}} a^{|S|} (\gamma(\nu-a^2))^{\frac{l-|S|}{2}} \mathbb{E}\left[\prod_{s\in S} z_{i_s j_s} z_{i_{s+1} j_s}\right] \mathbb{E}\left[\prod_{s\notin S} w_{i_s i_{s+1}}\right].
\end{aligned}$$

In the fourth line above, we restricted the summation to $i_s \neq i_{s+1}\, \forall s \in S$, as $v_{ii} = 0$ for each $i = 1, \ldots, \tilde{p}$ by Definition 4.8.

Let us write $\prod_{s\in S} z_{i_s j_s} z_{i_{s+1} j_s} = \prod_{i=1}^{\tilde{p}} \prod_{j=1}^{\tilde{n}} z_{ij}^{c_{ij}}$ where c_{ij} is the number of times z_{ij} appears in this product, and let us write $\prod_{s\notin S} w_{i_s i_{s+1}} = \prod_{i=1}^{\tilde{p}} w_{ii}^{a_{ii}} \prod_{i=1}^{\tilde{p}-1} \prod_{i'=i+1}^{\tilde{p}} w_{ii'}^{a_{ii'}} w_{i'i}^{b_{ii'}}$, where $a_{ii'}$ and $b_{ii'}$ are the numbers of times $w_{ii'}$ and $w_{i'i}$ appear in this product, respectively. Recall $\{z_{ij} : 1 \leq i \leq \tilde{p}, 1 \leq j \leq \tilde{n}\} \overset{iid}{\sim} \mathcal{N}(0,1)$, so $\mathbb{E}[\prod_{i=1}^{\tilde{p}} \prod_{j=1}^{\tilde{n}} z_{ij}^{c_{ij}}] \neq 0$ only if each c_{ij} is even (possibly zero), in which case this quantity is at least 1. Similarly, recall that $\{w_{ii} : 1 \leq i \leq \tilde{p}\} \cup \{\sqrt{2}\operatorname{Re} w_{ii'}, \sqrt{2}\operatorname{Im} w_{ii'} : 1 \leq i < i' \leq \tilde{p}\} \overset{iid}{\sim} \mathcal{N}(0,1)$, and $w_{ii'} = \overline{w_{i'i}}$ for $i > i'$. If $w = re^{i\theta}$ such that $\sqrt{2}\operatorname{Re} w, \sqrt{2}\operatorname{Im} w \overset{iid}{\sim} \mathcal{N}(0,1)$, then r and θ are independent with $r^2 \sim \chi_2^2/2$ and $\theta \sim \operatorname{Unif}[0, 2\pi)$. Then $\mathbb{E}[w^a \overline{w}^b] = \mathbb{E}[r^{a+b}]\mathbb{E}[e^{i(a-b)\theta}]$ for all nonnegative integers a, b, and this is 0 if $a \neq b$ and at least 1 if $a = b \geq 0$. Hence $\mathbb{E}[\prod_{i=1}^{\tilde{p}} w_{ii}^{a_{ii}} \prod_{i=1}^{\tilde{p}-1} \prod_{i'=i+1}^{\tilde{p}} w_{ii'}^{a_{ii'}} w_{i'i}^{b_{ii'}}] = 0$ unless $a_{ii'} = b_{ii'}$ for each $i' > i$ and a_{ii} is even (possibly zero) for each $1 \leq i \leq \tilde{p}$, in which case this quantity is also at least 1.

The above arguments imply, in particular, that $\mathbb{E}\left[\prod_{s\notin S} w_{i_s i_{s+1}}\right] = 0$ unless $l - |S|$ is even. As l is even by assumption, $l - |S|$ is even if and only if $|S|$ is also even, in which case $a^{|S|} = |a|^{|S|} \geq 0$. Hence each term of the sum in the above expression for $\mathbb{E}[\operatorname{Tr} M_{\tilde{n},\tilde{p}}^l]$ is nonnegative, so a lower bound

is obtained if we further restrict the summation to $i_s \neq i_{s+1}\, \forall s \in \{1,\dots,l\}$, i.e.

$$\mathbb{E}\left[\operatorname{Tr} M_{\tilde{n},\tilde{p}}^l\right] \geq \sum_{S\subseteq\{1,\dots,l\}} \sum_{\substack{i_1,\dots,i_l=1 \\ i_1\neq i_2, i_2\neq i_3,\dots,i_s\neq i_1}}^{\tilde{p}} \sum_{(j_s:s\in S)\in\{1,\dots,\tilde{n}\}^{|S|}} \tilde{n}^{-\frac{l+|S|}{2}} \left(\frac{\tilde{p}}{\tilde{n}}\right)^{-\frac{l-|S|}{2}} |a|^{|S|}(\gamma(\nu-a^2))^{\frac{l-|S|}{2}} \mathbb{E}\left[\prod_{s\in S} z_{i_s j_s} z_{i_{s+1} j_s}\right] \mathbb{E}\left[\prod_{s\notin S} w_{i_s i_{s+1}}\right].$$

We identify the combination of these three sums, over the nonzero terms, as a sum over all $(\tilde{p},\tilde{n})$-simple-labelings of an l-graph. Here, the first sum over S is over all choices of the subset of the n-vertices that have non-empty label. The second sum over $i_1,\dots,i_l$ is over all choices of p-labels, with condition (1) in Definition 4.10 that any two consecutive p-vertices have distinct labels being imposed by the constraints $i_1 \neq i_2, i_2 \neq i_3, \dots, i_l \neq i_1$ in the sum. The last sum over $(j_s : s \in S)$ is over all choices of n-labels for the n-vertices that have nonempty label. The product expression $\prod_{s\in S} z_{i_s j_s} z_{i_{s+1} j_s}$ then corresponds to a product, over all n-vertices with non-empty label and both p-vertices immediately preceding and following that n-vertex, of z_{ij}, where $j \in \{1,\dots,\tilde{n}\}$ is the n-label of the n-vertex and $i \in \{1,\dots,\tilde{n}\}$ is the p-label of the p-vertex. Similarly, the product expression $\prod_{s\notin S} w_{i_s i_{s+1}}$ corresponds to a product, over all n-vertices with empty label, of $w_{ii'}$, where i and i' are the p-labels of the p-vertices immediately preceding and immediately following this n-vertex, respectively. The condition that z_{ij} for each distinct pair (i,j) appears an even number of times, so that $\mathbb{E}[\prod_{s\in S} z_{i_s j_s} z_{i_{s+1} j_s}] \neq 0$, is precisely condition (2) in Definition 4.10, and the condition that each $w_{ii'}$ appears the same number of times as $w_{i'i}$, so that $\mathbb{E}[\prod_{s\notin S} w_{i_s i_{s+1}}] \neq 0$, is precisely condition (3) in Definition 4.10. As $\mathbb{E}[\prod_{s\in S} z_{i_s j_s} z_{i_{s+1} j_s}]\mathbb{E}[\prod_{s\notin S} w_{i_s i_{s+1}}] \geq 1$ whenever this quantity is nonzero, this implies

$$\mathbb{E}\left[\operatorname{Tr} M_{\tilde{n},\tilde{p}}^l\right] \geq \sum_{l\text{-graph } (\tilde{p},\tilde{n})\text{-simple-labelings}} \tilde{n}^{-\frac{l+\tilde{k}}{2}} \left(\frac{\tilde{p}}{\tilde{n}}\right)^{-\frac{l-\tilde{k}}{2}} |a|^{\tilde{k}}(\gamma(\nu-a^2))^{\frac{l-\tilde{k}}{2}},$$

where $\tilde{k} = |S|$ is the number of n-vertices in the simple-labeling with non-empty label. Any simple labeling with $\tilde{r}$ distinct p-labels and at most $\tilde{m}-\tilde{r}$ distinct non-empty n-labels has at most $\frac{\tilde{n}!}{(\tilde{n}-\tilde{m}+\tilde{r})!}\frac{\tilde{p}!}{(\tilde{p}-\tilde{r})!} \geq \tilde{n}^{\tilde{m}}\left(\frac{\tilde{p}}{\tilde{n}}\right)^{\tilde{r}}\left(\frac{\tilde{p}-l}{\tilde{p}}\right)^{l}\left(\frac{\tilde{n}-l}{\tilde{p}}\right)^{l}$ labelings in its equivalence class (where we have used $\tilde{m}-\tilde{r} \leq l$ and $\tilde{r} \leq l$). The desired result then follows upon identifying $\tilde{n}^{1-\tilde{\Delta}} = \tilde{n}^{-\frac{l+\tilde{k}}{2}+\tilde{m}}$. □

With Lemmas 4.7 and 4.14 established, the remainder of our proof of Proposition 4.1 involves a comparison of the upper bound in eq. (6) of Lemma 4.7 and the lower bound in eq. (7) of Lemma 4.14. The intuition for the comparison is the following: From Lemmas 4.4 and 4.11, we know that the excesses satisfy $\Delta(\mathcal{L}) \geq 0$ and $\tilde{\Delta}(\mathcal{L}) \geq 0$ in eqs. (6) and (7), respectively. Provided that the number of labelings with excesses $\Delta(\mathcal{L})$ and $\tilde{\Delta}(\mathcal{L})$ do not increase too rapidly as $\Delta(\mathcal{L})$ and $\tilde{\Delta}(\mathcal{L})$ increase, we expect that when n is large, the dominant contributions to the sums in eqs. (6) and (7) come from the labelings with zero excess. It may be shown that if we take any multi-labeling equivalence class $\mathcal{L}$ with excess $\Delta(\mathcal{L}) = 0$ and replace the labels of any n-vertex having more than one n-label with the empty label, then this mapping yields a valid simple-labeling equivalence class $\tilde{\mathcal{L}}$ with excess $\tilde{\Delta}(\tilde{\mathcal{L}}) = 0$, and furthermore, $\sum_{\mathcal{L}:\mathcal{L}\text{ maps to }\tilde{\mathcal{L}}} \prod_{s=1}^{l} \frac{a_{d_s}(\mathcal{L})}{(d_s(\mathcal{L})!)^{1/2}} = |a|^{|\tilde{k}(\tilde{\mathcal{L}})|}(\nu-a^2)^{\frac{l-\tilde{k}(\tilde{\mathcal{L}})}{2}}$. Hence, this mapping yields a direct correspondence between terms in eq. (6) with excess $\Delta(\mathcal{L}) = 0$ and terms in eq. (7) with excess $\tilde{\Delta}(\mathcal{L}) = 0$. To handle the terms in eq. (6) where $\Delta(\mathcal{L}) \neq 0$, we will extend this mapping to multi-labeling equivalence classes $\mathcal{L}$ with positive excess. We do this in the case $a \neq 0$, and the properties of this mapping that we will need are summarized in the following proposition.

Proposition 4.15. *Suppose $a \neq 0$ and $l \geq 2$. Let $\mathcal{C}$ denote the set of all multi-labeling equivalence classes of an l-graph, and let $\tilde{\mathcal{C}}$ denote the set of all simple-labeling equivalence classes of an l-graph. For $\mathcal{L} \in \mathcal{C}$, let $\Delta(\mathcal{L})$ be its excess and $r(\mathcal{L})$ be the number of distinct p-labels, and for $\tilde{\mathcal{L}} \in \tilde{\mathcal{C}}$, let $\tilde{\Delta}(\mathcal{L})$ be its excess, $\tilde{r}(\tilde{\mathcal{L}})$ be the number of distinct p-labels, and $\tilde{k}(\tilde{\mathcal{L}})$ be the number of n-vertices with non-empty label. Then there exists a map $\varphi : \mathcal{C} \to \tilde{\mathcal{C}}$ such that, for some constants $C_1, C_2, C_3, C_4 > 0$ depending only on D,*

1. *For all $\mathcal{L} \in \mathcal{C}$, $r(\mathcal{L}) = \tilde{r}(\tilde{\mathcal{L}})$,*
2. *For all $\mathcal{L} \in \mathcal{C}$, $\tilde{\Delta}(\varphi(\mathcal{L})) \leq C_1 \Delta(\mathcal{L})$, and*
3. *For any $\tilde{\mathcal{L}} \in \tilde{\mathcal{C}}$ and $\Delta_0 \geq 0$,*

$$\sum_{\substack{\mathcal{L} \in \varphi^{-1}(\tilde{\mathcal{L}}) \\ \Delta(\mathcal{L}) = \Delta_0}} \prod_{s=1}^{l} \frac{|a_{d_s(\mathcal{L})}|}{(d_s(\mathcal{L})!)^{1/2}} \leq \left(\frac{\sqrt{\nu}}{|a|}\right)^{C_2 \Delta_0} |a|^{\tilde{k}(\tilde{\mathcal{L}})} (\nu - a^2)^{\frac{l - \tilde{k}(\tilde{\mathcal{L}})}{2}} l^{C_3 + C_4 \Delta_0}.$$

This proposition allows us to control all terms in the sum in eq. (6), including those with positive excess, by terms in the sum in eq. (7), thus bypassing the need to directly control how the number of terms in the sum for eq. (6) corresponding to each value of $\Delta(\mathcal{L})$ grows with $\Delta(\mathcal{L})$. The proof of this proposition and the explicit construction of the map φ require some detailed combinatorial arguments, which we defer to Appendix A. Using this result, we may complete the proof of Proposition 4.1 in the case $a \neq 0$.

Proof of Proposition 4.1 (Case $a \neq 0$). For any $\varepsilon > 0$ and even integer $l \geq 2$,

$$\mathbb{P}\left[\|Q_{n,p}(X)\| > (1+\varepsilon)\|\mu_{a,\nu,\gamma}\|\right] \leq \mathbb{P}\left[\operatorname{Tr} Q_{n,p}(X)^l > ((1+\varepsilon)\|\mu_{a,\nu,\gamma}\|)^l\right] \leq \frac{\mathbb{E}[\operatorname{Tr} Q_{n,p}(X)^l]}{(1+\varepsilon)^l \|\mu_{a,\nu,\gamma}\|^l}.$$

By Lemma 4.7, Definition 4.5, and Proposition 4.15,

$$\begin{aligned}
\mathbb{E}[\operatorname{Tr} Q_{n,p}(X)^l] &\leq n \sum_{\mathcal{L} \in \mathcal{C}} \left(\frac{(12(\frac{l+Dl}{2}+1))^{12\alpha}}{n}\right)^{\Delta(\mathcal{L})} \left(\frac{p}{n}\right)^{r(\mathcal{L})} \left(\prod_{s=1}^{l} \frac{|a_{d_s(\mathcal{L})}|}{(d_s(\mathcal{L})!)^{1/2}}\right) \\
&= n \sum_{\tilde{\mathcal{L}} \in \tilde{\mathcal{C}}} \sum_{\Delta_0 = \left\lceil \frac{\tilde{\Delta}(\tilde{\mathcal{L}})}{C_1} \right\rceil}^{\frac{l+Dl}{2}+1} \sum_{\substack{\mathcal{L} \in \varphi^{-1}(\tilde{\mathcal{L}}) \\ \Delta(\mathcal{L}) = \Delta_0}} \left(\frac{(12(\frac{l+Dl}{2}+1))^{12\alpha}}{n}\right)^{\Delta_0} \left(\frac{p}{n}\right)^{\tilde{r}(\tilde{\mathcal{L}})} \prod_{s=1}^{l} \frac{|a_{d_s(\mathcal{L})}|}{(d_s(\mathcal{L})!)^{1/2}} \\
&\leq n \sum_{\tilde{\mathcal{L}} \in \tilde{\mathcal{C}}} \left(\frac{p}{n}\right)^{\tilde{r}(\tilde{\mathcal{L}})} \sum_{\Delta_0 = \left\lceil \frac{\tilde{\Delta}(\tilde{\mathcal{L}})}{C_1} \right\rceil}^{\frac{l+Dl}{2}+1} \left(\frac{(12(\frac{l+Dl}{2}+1))^{12\alpha}}{n}\right)^{\Delta_0} \\
&\qquad \left(\frac{\sqrt{\nu}}{|a|}\right)^{C_2 \Delta_0} |a|^{\tilde{k}(\tilde{\mathcal{L}})} (\nu - a^2)^{\frac{l - \tilde{k}(\tilde{\mathcal{L}})}{2}} l^{C_3 + C_4 \Delta_0} \\
&\leq n l^{C_3} \left(\tfrac{l+Dl}{2} + 2\right) \sum_{\tilde{\mathcal{L}} \in \tilde{\mathcal{C}}} \left(\frac{p}{n}\right)^{\tilde{r}(\tilde{\mathcal{L}})} |a|^{\tilde{k}(\tilde{\mathcal{L}})} (\nu - a^2)^{\frac{l - \tilde{k}(\tilde{\mathcal{L}})}{2}} \\
&\qquad \left(\frac{\left(12(\frac{l+Dl}{2}+1)\right)^{12\alpha} \left(\frac{\sqrt{\nu}}{|a|}\right)^{C_2} l^{C_4}}{n}\right)^{\frac{\tilde{\Delta}(\tilde{\mathcal{L}})}{C_1}},
\end{aligned}$$

where the last line holds for all sufficiently large n if $l \sim M \log n$, for any $M > 0$. Let

$$\tilde{n} = \left\lfloor \frac{n^{\frac{1}{C_1}}}{\left(12(\frac{l+lD}{2}+1)\right)^{\frac{12\alpha}{C_1}} \left(\frac{\sqrt{\nu}}{|a|}\right)^{\frac{C_2}{C_1}} l^{\frac{C_4}{C_1}}} \right\rfloor,$$

and let $\tilde{p} = \lfloor \frac{\tilde{n}p}{n} \rfloor$. Then $nl^{C_3}\left(\frac{l+Dl}{2}+2\right) \le n^2$ and $(\frac{p}{n})^{\tilde{r}(\tilde{\mathcal{L}})} \le (\frac{\tilde{p}}{\tilde{n}})^{\tilde{r}(\tilde{\mathcal{L}})}(1+\frac{\varepsilon}{4})^l$ for $l \sim M \log n$, any $M > 0$, and all sufficiently large n, so

$$\mathbb{E}[\operatorname{Tr} Q_{n,p}(X)^l] \le n^2 \left(1+\tfrac{\varepsilon}{4}\right)^l \sum_{\tilde{\mathcal{L}} \in \tilde{\mathcal{C}}} \left(\frac{1}{\tilde{n}}\right)^{\tilde{\Delta}(\tilde{\mathcal{L}})} \left(\frac{\tilde{p}}{\tilde{n}}\right)^{\tilde{r}(\tilde{\mathcal{L}})} |a|^{\tilde{k}(\tilde{\mathcal{L}})} (\nu - a^2)^{\frac{l-\tilde{k}(\tilde{\mathcal{L}})}{2}}.$$

On the other hand, by Lemma 4.14,

$$\left(1-\tfrac{\varepsilon}{4}\right)^l \tilde{n} \sum_{\tilde{\mathcal{L}} \in \tilde{\mathcal{C}}} \left(\frac{1}{\tilde{n}}\right)^{\tilde{\Delta}(\tilde{\mathcal{L}})} \left(\frac{\tilde{p}}{\tilde{n}}\right)^{\tilde{r}(\tilde{\mathcal{L}})} |a|^{\tilde{k}(\tilde{\mathcal{L}})} (\nu - a^2)^{\frac{l-\tilde{k}(\tilde{\mathcal{L}})}{2}} \le \mathbb{E}\left[\operatorname{Tr} M_{\tilde{n},\tilde{p}}^l\right]$$

for all sufficiently large n. Since $\frac{\tilde{p}}{\tilde{n}} \to \gamma$ and $l \sim MC_1 \log \tilde{n}$ if $l \sim M \log n$, Proposition 4.9 implies $\mathbb{E}[\operatorname{Tr} M_{\tilde{n},\tilde{p}}^l] \le \tilde{p}\, \mathbb{E}[\|M_{\tilde{n},\tilde{p}}\|^l] \le \tilde{p}\left(\|\mu_{a,\nu,\gamma}\|\left(1+\frac{\varepsilon}{4}\right)\right)^l$ for all large n. Thus

$$\mathbb{P}\left[\|Q_{n,p}(X)\| > (1+\varepsilon)\|\mu_{a,\nu,\gamma}\|\right] \le n^2 \frac{\tilde{p}}{\tilde{n}} \left(\frac{\left(1+\frac{\varepsilon}{4}\right)^2}{\left(1-\frac{\varepsilon}{4}\right)(1+\varepsilon)}\right)^l.$$

Taking $l \sim M \log n$ with $M > 0$ sufficiently large such that $M \log \frac{(1+\frac{\varepsilon}{4})^2}{(1-\frac{\varepsilon}{4})(1+\varepsilon)} < -4$ (which is possible for any sufficiently small $\varepsilon > 0$), this implies $\mathbb{P}\left[\|Q_{n,p}(X)\| > (1+\varepsilon)\|\mu_{a,\nu,\gamma}\|\right] < \frac{1}{n^2}$ for all large n. Then $\limsup_{n,p\to\infty} \|Q_{n,p}(X)\| \le (1+\varepsilon)\|\mu_{a,\nu,\gamma}\|$ a.s., by the Borel-Cantelli lemma. This holds for all sufficiently small $\varepsilon > 0$, so the proposition follows. □

Proposition 4.1 in the case $a = 0$ may be easily established from the $a \neq 0$ case via the following continuity argument.

Lemma 4.16. *For the measure $\mu_{a,\nu,\gamma}$ in Definition 2.3, $\|\mu_{a,\nu,\gamma}\|$ is continuous in (a,ν,γ).*

Proof. Let $W_{\tilde{p}}$ and $V_{\tilde{n},\tilde{p}}$ be as in Definition 4.8, and let $M_{\tilde{n},\tilde{p}}(a,\nu,\gamma) = \sqrt{\frac{\gamma(\nu-a^2)}{\tilde{p}}} W_{\tilde{p}} + \frac{a}{\tilde{n}} V_{\tilde{n},\tilde{p}}$. For any fixed (a,ν,γ) and $(a_i,\nu_i,\gamma_i)_{i=1}^\infty$ such that $\lim_{i\to\infty}(a_i,\nu_i,\gamma_i) = (a,\nu,\gamma)$, by Lemma B.1, $\|M_{\tilde{n},\tilde{p}}(a,\nu,\gamma)\| \to \|\mu_{a,\nu,\gamma}\|$, $\|M_{\tilde{n},\tilde{p}}(a_i,\nu_i,\gamma_i)\| \to \|\mu_{a_i,\nu_i,\gamma_i}\|$ for each i, and $\limsup_{n,p\to\infty} \frac{1}{\sqrt{\tilde{p}}}\|W_{\tilde{p}}\| < \infty$ and $\limsup_{n,p\to\infty} \frac{1}{\tilde{n}}\|V_{\tilde{n},\tilde{p}}\| < \infty$ on an event having probability 1. Then on this event, for each i,

$$\begin{aligned}
\left|\|\mu_{a_i,\nu_i,\gamma_i}\| - \|\mu_{a,\nu,\gamma}\|\right| &\le \limsup_{n,p\to\infty} \left|\|M_{\tilde{n},\tilde{p}}(a_i,\nu_i,\gamma_i)\| - \|M_{\tilde{n},\tilde{p}}(a,\nu,\gamma)\|\right| \\
&\le \limsup_{n,p\to\infty} \|M_{\tilde{n},\tilde{p}}(a_i,\nu_i,\gamma_i) - M_{\tilde{n},\tilde{p}}(a,\nu,\gamma)\| \\
&\le \left|\sqrt{\gamma_i(\nu_i - a_i^2)} - \sqrt{\gamma(\nu-a^2)}\right| \limsup_{n,p\to\infty} \frac{1}{\sqrt{\tilde{p}}}\|W_{\tilde{p}}\| + |a - a_i| \limsup_{n,p\to\infty} \frac{1}{\tilde{n}}\|V_{\tilde{n},\tilde{p}}\|,
\end{aligned}$$

which implies $\lim_{i\to\infty} \|\mu_{a_i,\nu_i,\gamma_i}\| = \|\mu_{a,\nu,\gamma}\|$. □

Proof of Proposition 4.1 (Case $a = 0$). Suppose $k(x)$ is a polynomial function such that the coefficient of the linear term in its Hermite polynomial decomposition is zero. For any $a > 0$, let $k_a(x) = k(x) + ax$, and let $Q_{n,p,a}(X)$ be the matrix as defined in Definition 3.2 for the kernel function k_a. Then $Q_{n,p,a}(X) = Q_{n,p}(X) + \frac{a}{n} V_{n,p}(X)$, where $V_{n,p}(X)$ has zero diagonal and

equals XX^T off of the diagonal. By Proposition 4.1 for the $a \neq 0$ case, established above, $\limsup_{n,p\to\infty} \|Q_{n,p,a}(X)\| \leq \|\mu_{a,(\nu+a^2),\gamma}\|$. As a consequence of the main theorem in [16] and the observation $\lim_{n,p\to\infty} \sup_{i=1}^p \left(\frac{\|X_i\|^2}{n} - 1\right) = 0$, which follows easily under the Assumption 2.1, we have $\limsup_{n,p\to\infty} \|\frac{1}{n}V_{n,p}(X)\| \leq C_\gamma$ for some constant $C_\gamma > 0$. This implies $\limsup_{n,p\to\infty} \|Q_{n,p}(X)\| \leq \|\mu_{a,(\nu+a^2),\gamma}\| - aC_\gamma$ for any $a > 0$, and the desired result follows from Lemma 4.16 upon taking $a \to 0$. □

5. Analyzing the remainder matrices $R_{n,p}(X)$ and $S_{n,p}(X)$

To conclude the proof of Theorem 2.6, we analyze in this section the remainder matrices $R_{n,p}(X)$ and $S_{n,p}(X)$ of Definition 3.2.

Lemma 5.1. *As $n, p \to \infty$ with $\frac{p}{n} \to \gamma \in (0,\infty)$, $\lim_{n,p\to\infty} \|S_{n,p}(X)\| = 0$ almost surely.*

Proof. Note that $\|S_{n,p}(X)\| \leq \|S_{n,p}(X)\|_F \leq p \max_{1\leq i,i'\leq p} |s_{ii'}|$ where $\|\cdot\|_F$ is the Frobenius norm. By Definition 3.2 and Proposition 3.1, for any $1 \leq i, i' \leq p$ and $\alpha > 0$, $|s_{ii'}| \leq n^{-\frac{3}{2}+\alpha} \sum_{d=1}^D |a_d|$ with probability at least $1 - n^{-4}$, for all large n. Then by a union bound, $p \max_{1\leq i,i'\leq p} |s_{ii'}| \leq pn^{-\frac{3}{2}+\alpha}$ with probability at least $1 - p^2n^{-4}$. As $\frac{p}{n} \to \gamma \in (0,\infty)$, taking any $\alpha < \frac{1}{2}$ and $\varepsilon > 0$, this implies $\mathbb{P}[\|S_{n,p}(X)\| > \varepsilon] \leq p^2n^{-4}$ for all large n, so $\limsup_{n,p\to\infty} \|S_{n,p}(X)\| \leq \varepsilon$ a.s. by the Borel-Cantelli lemma. As $\varepsilon > 0$ is arbitrary, the result follows. □

Definition 5.2. *For $d \geq 2$, let $R_{n,p,d}(X) := (r_{ii'} : 1 \leq i, i' \leq p) \in \mathbb{R}^{p\times p}$, with entries*

$$r_{ii'} = \begin{cases} \frac{\binom{d}{2}}{\sqrt{d!}} n^{-\frac{d+1}{2}} \sum_{\substack{j_1,\ldots,j_{d-1}=1 \\ j_1\neq j_2\neq\cdots\neq j_{d-1}}}^n \left((x_{ij_1}^2 x_{i'j_1}^2 - 1) \prod_{a=2}^{d-1} x_{ij_a} x_{i'j_a} \right) & i \neq i' \\ 0 & i = i'. \end{cases}$$

Note that $R_{n,p}(X)$ in Definition 3.2 is given by $R_{n,p}(X) = \sum_{d=2}^D a_d R_{n,p,d}(X)$.

Lemma 5.3. *As $n, p \to \infty$ with $\frac{p}{n} \to \gamma \in (0,\infty)$, $\lim_{n,p\to\infty} \|R_{n,p,d}(X)\| = 0$ a.s. for any $d \geq 3$.*

Proof. Letting $i_7 := i_1$ for notational convenience, note that

$$\begin{aligned} \mathbb{E}\left[\operatorname{Tr} R_{n,p,d}(X)^6\right] &= \sum_{\substack{i_1,\ldots,i_6=1 \\ i_1\neq i_2, i_2\neq i_3,\ldots,i_6\neq i_1}}^p \mathbb{E}\left[\prod_{s=1}^6 r_{i_s i_{s+1}}\right] \\ &= \frac{\binom{d}{2}^6}{(d!)^3} n^{-3(d+1)} \sum_{\substack{i_1,\ldots,i_6=1 \\ i_1\neq i_2, i_2\neq i_3,\ldots,i_6\neq i_1}}^p \sum_{\substack{j_1^1,\ldots,j_{d-1}^1=1 \\ j_1^1\neq j_2^1\neq\cdots\neq j_{d-1}^1}}^n \cdots \sum_{\substack{j_1^6,\ldots,j_{d-1}^6=1 \\ j_1^6\neq j_2^6\neq\cdots\neq j_{d-1}^6}}^n \\ &\qquad \mathbb{E}\left[\prod_{s=1}^6 \left((x_{i_s j_1^s}^2 x_{i_{s+1} j_1^s}^2 - 1) \prod_{a=2}^{d-1} x_{i_s j_a^s} x_{i_{s+1} j_a^s} \right)\right]. \end{aligned}$$

Consider the term

$$\mathbb{E}\left[\prod_{s=1}^6 \left((x_{i_s j_1^s}^2 x_{i_{s+1} j_1^s}^2 - 1) \prod_{a=2}^{d-1} x_{i_s j_a^s} x_{i_{s+1} j_a^s} \right)\right]. \tag{8}$$

Suppose that there is some $j^* \in \{1,\ldots,n\}$ such that $j_a^s = j^*$ for exactly one pair of indices $(s,a) \in \{1,\ldots,6\} \times \{1,\ldots,d-1\}$. If $a = 1$, then the only term of the product in eq. (8) that contains a variable equal to x_{ij^*} for any i is $(x_{i_s j_1^s}^2 x_{i_{s+1} j_1^s}^2 - 1)$, and if $a \geq 2$, then the only two terms that contain a variable equal to x_{ij^*} for any i are $x_{i_s j_a^s}$ and $x_{i_{s+1} j_a^s}$. In either case, as $i_s \neq i_{s+1}$,

this implies eq. (8) is zero by independence of the entries of X. Hence, in order for eq. (8) to be nonzero, each distinct j^* that appears in the product must appear as j_a^s for at least two pairs (s,a).

Suppose, next, that there is some $i^* \in \{1,\dots,p\}$ such that $i_s = i^*$ for exactly one index $s \in \{1,\dots,6\}$. Consider the label $j^* = j_a^{s-1}$ for any $a \in \{2,\dots,d-1\}$ (identifying $s-1 \equiv 6$ if $s=1$). Such a label j^* exists when $d \geq 3$. If $j^* \neq j_{a'}^s$ for all $a' \in \{1,\dots,d-1\}$, then the variable $x_{i^*j^*}$ appears exactly once in the product in eq. (8), and so eq. (8) is zero. If $j^* = j_1^s$, then the variable $x_{i^*j^*}$ appears twice, once as the term $x_{i_s j_a^{s-1}}$ and once in the term $(x_{i_s j_1^s}^2 x_{i_{s+1} j_1^s}^2 - 1)$. The product of these terms is $x_{i^*j^*}^3 x_{i_{s+1}j^*} - x_{i^*j^*}$, and as $\mathbb{E}[x_{i^*j^*}^3] = 0$ and $\mathbb{E}[x_{i^*j^*}] = 0$, this implies eq. (8) is again zero. Hence, in order for eq. (8) to be nonzero, j_a^{s-1} must equal $j_{a'}^s$ for some $a' \geq 2$.

Consider first the case where there are at most 4 distinct values in $\{i_1,\dots,i_6\}$. Note that if eq. (8) is nonzero, then there are at most $\frac{6(d-1)}{2} = 3d-3$ distinct values of j_a^s in the product, by our previous argument. There are at most $p^4 n^{3d-3} C_d$ total choices of indices $(i_1,\dots,i_6)$ and $(j_a^s : 1 \leq a \leq d-1, 1 \leq s \leq 6)$ such that $|\{i_1,\dots,i_6\}| \leq 4$ and $|\{j_a^s : 1 \leq a \leq d-1, 1 \leq s \leq 6\}| \leq 3d-3$, where C_d is a combinatorial constant depending on d but not on n and p. Then

$$\sum_{\substack{i_1,\dots,i_6=1 \\ i_1 \neq i_2, i_2 \neq i_3, \dots, i_6 \neq i_1 \\ |\{i_1,\dots,i_6\}| \leq 4}}^{p} \sum_{\substack{j_1^1,\dots,j_{d-1}^1=1 \\ j_1^1 \neq \dots \neq j_{d-1}^1}}^{n} \cdots \sum_{\substack{j_1^6,\dots,j_{d-1}^6=1 \\ j_1^6 \neq \dots \neq j_{d-1}^6}}^{n} \mathbb{E}\left[\prod_{s=1}^{6}\left((x_{i_s j_1^s}^2 x_{i_{s+1} j_1^s}^2 - 1)\prod_{a=2}^{d-1} x_{i_s j_a^s} x_{i_{s+1} j_a^s}\right)\right] \leq C_{d,\alpha} p^4 n^{3d-3}$$

for a constant $C_{d,\alpha}$ depending on d and the value of α in Assumption 2.1.

If there are 5 distinct values in $\{i_1,\dots,i_6\}$, then either $i_s = i_{s+2}$ for some s or $i_s = i_{s+3}$ for some s (where $s+2$ and $s+3$ are taken modulo 6), with the remaining indices all distinct. Suppose without loss of generality that i_2 and i_3 are distinct from $\{i_1, i_4, i_5, i_6\}$. Then, letting $i^* := i_2$ and $j^* := j_2^1$, we note that i_2 is the unique index in $\{i_1,\dots,i_6\}$ equal to i^*, so for eq. (8) to be nonzero, we must have $j^* = j_a^2$ for some $a \geq 2$, by our previous argument. The same argument applied to $i^* := i_3$ then shows that we must have $j^* = j_{a'}^3$ for some $a' \geq 2$. Then there are at least three pairs (s,a) for which $j_a^s = j^*$. This implies that there are at most $3d-4$ distinct values of j_a^s (as if there were $3d-3$ distinct values and each value must equal j_a^s for at least two pairs (s,a), then no such value can equal j_a^s for more than two pairs). There are at most $p^5 n^{3d-4} C_d'$ total choices of indices $(i_1,\dots,i_6)$ and $(j_a^s : 1 \leq a \leq d-1, 1 \leq s \leq 6)$ such that $|\{i_1,\dots,i_6\}| = 5$ and $|\{j_a^s : 1 \leq a \leq d-1, 1 \leq s \leq 6\}| \leq 3d-4$, where C_d' is a combinatorial constant depending on d but not on n and p. Then

$$\sum_{\substack{i_1,\dots,i_6=1 \\ i_1 \neq i_2, i_2 \neq i_3, \dots, i_6 \neq i_1 \\ |\{i_1,\dots,i_6\}| = 5}}^{p} \sum_{\substack{j_1^1,\dots,j_{d-1}^1=1 \\ j_1^1 \neq \dots \neq j_{d-1}^1}}^{n} \cdots \sum_{\substack{j_1^6,\dots,j_{d-1}^6=1 \\ j_1^6 \neq \dots \neq j_{d-1}^6}}^{n} \mathbb{E}\left[\prod_{s=1}^{6}\left((x_{i_s j_1^s}^2 x_{i_{s+1} j_1^s}^2 - 1)\prod_{a=2}^{d-1} x_{i_s j_a^s} x_{i_{s+1} j_a^s}\right)\right] \leq C_{d,\alpha}' p^5 n^{3d-4}$$

for a constant $C_{d,\alpha}'$.

Finally, if $|\{i_1,\dots,i_6\}| = 6$, then letting $j^* = j_2^1$ and $i^* = i_2$, our previous argument implies that for eq. (8) to be nonzero, there must be some $a \geq 2$ such that $j^* = j_a^2$. Similarly, for each $s = 3,\dots,6$, there must be some $a \geq 2$ such that $j^* = j_a^s$. Then there are exactly 6 pairs (s,a) such that $j_a^s = j^*$, so the number of distinct values of j_a^s is at most $\frac{6(d-1)-6}{2} + 1 = 3d-5$. There are at most $p^6 n^{3d-5} C_d''$ total choices of indices $(i_1,\dots,i_6)$ and $(j_a^s : 1 \leq a \leq d-1, 1 \leq s \leq 6)$ such that $|\{i_1,\dots,i_6\}| = 6$ and $|\{j_a^s : 1 \leq a \leq d-1, 1 \leq s \leq 6\}| \leq 3d-5$, where C_d'' is a combinatorial

constant depending on d but not on n and p. Then

$$\sum_{\substack{i_1,\dots,i_6=1\\ i_1\neq i_2, i_2\neq i_3,\dots,i_6\neq i_1\\ |\{i_1,\dots,i_6\}|=6}}^{p} \sum_{\substack{j_1^1,\dots,j_{d-1}^1=1\\ j_1^1\neq\dots\neq j_{d-1}^1}}^{n} \cdots \sum_{\substack{j_1^6,\dots,j_{d-1}^6=1\\ j_1^6\neq\dots\neq j_{d-1}^6}}^{n} \mathbb{E}\left[\prod_{s=1}^{6}\left((x_{i_s j_1^s}^2 x_{i_{s+1} j_1^s}^2 - 1)\prod_{a=2}^{d-1} x_{i_s j_a^s} x_{i_{s+1} j_a^s}\right)\right] \le C''_{d,\alpha} p^6 n^{3d-5}$$

for a constant $C''_{d,\alpha}$. Putting this together,

$$\mathbb{E}\left[\operatorname{Tr} R_{n,p,d}(X)^6\right] \le \frac{C_{d,\alpha,\gamma}}{n^2}$$

for some constant $C_{d,\alpha,\gamma}$ depending also on γ. Then for any $\varepsilon > 0$,

$$\mathbb{P}\left[\|R_{n,p,d}(X)\| > \varepsilon\right] \le \frac{\mathbb{E}[\operatorname{Tr} R_{n,p,d}(X)^6]}{\varepsilon^6} \le \frac{C_{d,\alpha,\gamma}}{\varepsilon^6 n^2},$$

so the Borel-Cantelli lemma implies $\limsup_{n,p\to\infty} \|R_{n,p,d}(X)\| \le \varepsilon$ almost surely. As this holds for all $\varepsilon > 0$, the result follows. □

Lemma 5.4. *Let $R_{n,p,d}(X)$ be as in Definition 5.2, and let $\tilde{R}_{n,p}(X)$ be as in Theorem 2.6. As $n, p \to \infty$ with $\frac{p}{n} \to \gamma \in (0,\infty)$, $\lim_{n,p\to\infty} \|a_2 R_{n,p,2}(X) - \tilde{R}_{n,p}(X)\| \to 0$.*

Proof. Let $T_{n,p}(X) = a_2 R_{n,p,2}(X) - \tilde{R}_{n,p}(X)$. Then $T_{n,p}(X)$ has entries

$$\begin{aligned} t_{ii'} &= \begin{cases} \frac{a_2}{\sqrt{2}} n^{-\frac{3}{2}} \sum_{j=1}^n \left((x_{ij}^2 x_{i'j}^2 - 1) - (x_{ij}^2 - 1) - (x_{i'j}^2 - 1)\right) & i \neq i' \\ 0 & i = i' \end{cases} \\ &= \begin{cases} \frac{a_2}{\sqrt{2}} n^{-\frac{3}{2}} \sum_{j=1}^n (x_{ij}^2 - 1)(x_{i'j}^2 - 1) & i \neq i' \\ 0 & i = i' \end{cases}. \end{aligned}$$

Thus, excluding the diagonal, $T_{n,p}(X)$ equals $\frac{a_2}{\sqrt{2}} n^{-\frac{3}{2}} YY^T$ where $Y = (y_{ij}) \in \mathbb{R}^{p\times n}$ and $y_{ij} = x_{ij}^2 - 1$. By Assumption 2.1 and the main theorem of [16], $\frac{1}{n}\|YY^T\|$ converges to a finite limit almost surely, so $n^{-\frac{3}{2}}\|YY^T\| \to 0$. For any $\varepsilon > 0$ and any i,

$$\mathbb{P}\left[n^{-\frac{3}{2}} \sum_{j=1}^n y_{ij}^2 - \mathbb{E}[y_{ij}^2] \ge \varepsilon\right] \le \frac{\mathbb{E}\left[\left(\sum_{j=1}^n y_{ij}^2 - \mathbb{E}[y_{ij}^2]\right)^4\right]}{\varepsilon^4 n^6} \le \frac{C(\varepsilon,\alpha)}{n^4}$$

for some constant $C(\varepsilon,\alpha) > 0$ depending on ε and α in Assumption 2.1. Then

$$\mathbb{P}\left[\max_{i=1}^p n^{-\frac{3}{2}} \sum_{j=1}^n y_{ij}^2 \ge \varepsilon\right] \le \frac{pC(\varepsilon,\alpha)}{n^4} \le \frac{C(\varepsilon,\alpha,\gamma)}{n^3}$$

for all large n, and hence $\left\|T_{n,p}(X) - \frac{a_2}{\sqrt{2}} n^{-\frac{3}{2}} YY^T\right\| \to 0$ almost surely by the Borel-Cantelli lemma, implying $\|T_{n,p}(X)\| \to 0$ almost surely. □

We may now conclude the proof of Theorem 2.6.

Proof of Theorem 2.6. By Remark 3.3, $K_{n,p}(X) = Q_{n,p}(X) + R_{n,p}(X) + S_{n,p}(X)$. As $R_{n,p}(X) = \sum_{d=2}^D a_d R_{n,p,d}(X)$, where $R_{n,p,d}(X)$ is as in Definition 5.2, Proposition 4.1 and Lemmas 5.1, 5.3, and 5.4 imply $\limsup_{n,p\to\infty} \|K_{n,p}(X) - \tilde{R}_{n,p}(X)\| \le \|\mu_{a,\nu,\gamma}\|$. On the other hand, Theorem 2.5 implies, for any $\varepsilon > 0$, $\lim_{n,p\to\infty} \frac{1}{p}\sum_{i=1}^p \mathbb{1}\{|\lambda_i(K_{n,p}(X))| \in [\|\mu_{a,\nu,\gamma}\| - \varepsilon, \|\mu_{a,\nu,\gamma}\|]\} > cp$ for some constant $c := c(\varepsilon) > 0$ a.s., so in particular, $\liminf_{n,p\to\infty} \max(\lambda_3(K_{n,p}(X)), -\lambda_{p-2}(K_{n,p}(X))) \ge$

$\|\mu_{a,\nu,\gamma}\|$ a.s., where $\lambda_3(K_{n,p}(X))$ and $\lambda_{p-2}(K_{n,p}(X))$ are the third-largest and third-smallest eigenvalues of $K_{n,p}(X)$. As $\tilde{R}_{n,p}(X)$ has rank two, Weyl's eigenvalue inequality implies $\lambda_3(K_{n,p}(X)) \leq \lambda_{\max}(K_{n,p}(X) - \tilde{R}_{n,p}(X))$ and $-\lambda_{p-2}(K_{n,p}(X)) \leq -\lambda_{\min}(K_{n,p}(X) - \tilde{R}_{n,p}(X))$, hence establishing $\liminf_{n,p\to\infty} \|K_{n,p}(X) - \tilde{R}_{n,p}(X)\| \geq \|\mu_{a,\nu,\gamma}\|$ almost surely. This concludes the proof of the theorem upon setting $\tilde{K}_{n,p}(X) = K_{n,p}(X) - \tilde{R}_{n,p}(X)$. □

6. Extension to odd kernel functions for Gaussian observations

In this section, we prove Theorem 2.10. Our proof will rely on the following two results, the first giving a polynomial approximation of the kernel function k and the second providing a general concentration inequality that will allow us to handle the remainder term from this polynomial approximation.

Theorem 6.1 ([8])**.** *Suppose $w(x)$ is an even, lower semi-continuous function on $\mathbb{R}$ with $1 \leq w(x) < \infty$, such that $\log w(x)$ is a convex function of $\log x$. Let C_w be the class of continuous functions on $\mathbb{R}$ such that $\lim_{|x|\to\infty} \frac{f(x)}{w(x)} = 0$ for all $f \in C_w$, and suppose C_w contains all polynomial functions. If $\int_1^\infty \frac{\log w(x)}{x^2} dx = \infty$, then for any $f \in C_w$ and $\varepsilon > 0$, there exists a polynomial P such that $|f(x) - P(x)| < \varepsilon w(x)$ for all $x \in \mathbb{R}$.*

Proof. See the first Theorem on page 956 of [8]. □

Proposition 6.2. *Let k be an odd and differentiable function with $|k'(x)| \leq e^{\beta|x|}$ for some $\beta > 0$ and all $x \in \mathbb{R}$. Let $X \in \mathbb{R}^{p\times n}$ have entries $x_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$, and let $K_{n,p}(X)$ be the kernel inner-product matrix with kernel k as in Definition 2.2. Suppose $p, n \to \infty$ with $\frac{p}{n} \to \gamma$. Then there exist constants $N_{\beta,\gamma}, C_{\beta,\gamma} > 0$, depending only on β and γ, such that*

$$\mathbb{P}[\|K_{n,p}(X)\| > C_{\beta,\gamma}] \leq \frac{C_{\beta,\gamma}}{n^2}$$

for all $n \geq N_{\beta,\gamma}$.

Assuming the above Proposition, let us first prove Theorem 2.10.

Proof of Theorem 2.10. By the given conditions, there exists $\beta > 0$ such that $\lim_{|x|\to\infty} \frac{|k'(x)|}{e^{\beta|x|}} = 0$. Applying Theorem 6.1 with $w(x) = e^{\beta|x|}$, for any $\varepsilon > 0$, there exists a polynomial $\dot{q}$ such that $|k'(x) - \dot{q}(x)| < \varepsilon e^{\beta|x|}$ for all $x \in \mathbb{R}$. As k is an odd function, k' is even, so we may take $\dot{q}$ to be an even polynomial function. (Otherwise, take the polynomial to be $\frac{1}{2}(\dot{q}(x) + \dot{q}(-x))$.) Let $q(x) = \int_0^x \dot{q}(x)dx$ for all $x \in \mathbb{R}$, and let $r(x) = k(x) - q(x)$. Then q is an odd polynomial function, r is hence also an odd function, and $|r'(x)| < \varepsilon e^{\beta|x|}$ by construction. Let $Q_{n,p}(X)$ be the kernel inner-product matrix with kernel function q as in Definition 2.2, and let $R_{n,p}(X)$ be the kernel inner-product matrix with kernel function $r(x)$, so that $K_{n,p}(X) = Q_{n,p}(X) + R_{n,p}(X)$. By Proposition 6.2, there exists a constant $C_{\beta,\gamma} > 0$ such that $\mathbb{P}[\|R_{n,p}(X)\| > \varepsilon C_{\beta,\gamma}] < \frac{C_{\beta,\gamma}}{n^2}$, for all sufficiently large n, so $\limsup_{n,p\to\infty} \|R_{n,p}(X)\| \leq \varepsilon C_{\beta,\gamma}$ almost surely. On the other hand, if $q(x) = a_{0,\varepsilon} + a_{1,\varepsilon}h_1(x) + \ldots + a_{D,\varepsilon}h_D(x)$ where $h_1, \ldots, h_D$ are the orthonormal Hermite polynomials as in Definition 2.4, then $a_{j,\varepsilon} = 0$ for all even j since q is an odd function, and hence, in particular, $a_2 = 0$ and $\mathbb{E}[q(\xi)] = 0$ for $\xi \sim \mathcal{N}(0,1)$. Then by Theorem 2.6, $\|Q_{n,p}(X)\| \to \|\mu_{a_\varepsilon,\nu_\varepsilon,\gamma}\|$ where $a_\varepsilon = a_{1,\varepsilon}$ and $\nu_\varepsilon = \sum_{d=1}^D a_{d,\varepsilon}^2$. Hence

$$\|\mu_{a_\varepsilon,\nu_\varepsilon,\gamma}\| - \varepsilon C_{\beta,\gamma} \leq \liminf_{n,p\to\infty} \|K_{n,p}(X)\| \leq \limsup_{n,p\to\infty} \|K_{n,p}(X)\| \leq \|\mu_{a_\varepsilon,\nu_\varepsilon,\gamma}\| + \varepsilon C_{\beta,\gamma}$$

for any $\varepsilon > 0$. Note that $|k(x) - q(x)| \leq \frac{\varepsilon}{\beta} e^{\beta|x|}$ for all $x \in \mathbb{R}$, so by the dominated convergence theorem, $\lim_{\varepsilon\to 0} \mathbb{E}[(k(\xi) - q(\xi))^2] = 0$ for $\xi \sim \mathcal{N}(0,1)$. Then $a_\varepsilon \to a$ and $\nu_\varepsilon \to \nu$ as $\varepsilon \to 0$, where

a and ν are defined as in Theorem 2.5 for the kernel function k. By Lemma 4.16, this implies $\lim_{\varepsilon\to 0}\|\mu_{a_\varepsilon,\nu_\varepsilon,\gamma}\| \to \|\mu_{a,\nu,\gamma}\|$, and hence $\lim_{n,p\to\infty}\|K_{n,p}(X)\| = \|\mu_{a,\nu,\gamma}\|$. □

In the remainder of this section, we prove Proposition 6.2. Our proof relies on the covering bound $\|K_{n,p}(X)\| = \sup_{y\in\mathbb{R}^p:\|y\|\le 1} y^T K_{n,p}(X) y \le C \sup_{y\in D_2^p} y^T K_{n,p}(X) y$ for an appropriate choice of covering set $D_2^p \subset \{y\in\mathbb{R}^p : \|y\|\le 1\}$ and sufficiently large constant C. The specific construction of D_2^p and method of bounding $\sup_{y\in D_2^p} y^T K_{n,p}(X) y$ are inspired by a similar argument in [19].

Definition 6.3. *Let $\mathcal{G}(\beta) \subset \mathbb{R}^{p\times n}\times\mathbb{R}^{p\times n}$ be the set of pairs (X, X') of $p\times n$ matrices such that $\|X\| \le \sqrt{p}+2\sqrt{n}$, $\|X'\| \le \sqrt{p}+2\sqrt{n}$, and for all $l = 1,\dots,p$,*

$$\frac{1}{p}\sum_{\substack{i=1\\ i\neq l}}^{p} \exp\left(\frac{16\beta}{\sqrt{n}}|X_i^T X_l|\right) \le 3e^{256\beta^2},$$

$$\frac{1}{p}\sum_{\substack{i=1\\ i\neq l}}^{p} \exp\left(\frac{16\beta}{\sqrt{n}}|X_i'^T X_l|\right) \le 3e^{256\beta^2},$$

$$\frac{1}{p}\sum_{\substack{i=1\\ i\neq l}}^{p} \exp\left(\frac{16\beta}{\sqrt{n}}|X_i^T X_l'|\right) \le 3e^{256\beta^2},$$

$$\frac{1}{p}\sum_{\substack{i=1\\ i\neq l}}^{p} \exp\left(\frac{16\beta}{\sqrt{n}}|X_i'^T X_l'|\right) \le 3e^{256\beta^2},$$

Lemma 6.4. *Let $X, X' \in \mathbb{R}^{p\times n}$ with $x_{ij}, x'_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$. Then for all sufficiently large p (i.e. $p > p_0(\beta)$),*

$$\mathbb{P}[(X,X') \notin \mathcal{G}(\beta)] \le 4e^{-\frac{n}{2}} + \frac{484e^{384\beta^2}}{p^2} + 4pe^{-\frac{n}{8}}.$$

Proof. By Corollary 5.35 of [31], $\mathbb{P}\left[\|X\| > \sqrt{p}+2\sqrt{n}\right] \le 2e^{-\frac{n}{2}}$, and similarly for X'.

For $\xi\sim\mathcal{N}(0,1)$ and $c>0$, $\mathbb{E}[e^{c|\xi|}] \le \mathbb{E}[e^{c\xi}]+\mathbb{E}[e^{-c\xi}] = 2e^{\frac{c^2}{2}}$, and $\mathrm{Var}[e^{c|\xi|}] \le \mathbb{E}[e^{2c|\xi|}] \le 2e^{2c^2}$. Then for $\xi_1,\dots,\xi_p \overset{iid}{\sim}\mathcal{N}(0,1)$, denoting $f(\xi) = e^{c|\xi|} - \mathbb{E}[e^{c|\xi|}]$,

$$\begin{aligned}\mathbb{P}\left[\frac{1}{p}\sum_{i=1}^p e^{c|\xi_i|} > 3e^{\frac{c^2}{2}}\right] &\le \mathbb{P}\left[\frac{1}{p}\sum_{i=1}^p f(\xi_i) > e^{\frac{c^2}{2}}\right] \le \frac{e^{-3c^2}}{p^6}\mathbb{E}\left[\left(\sum_{i=1}^p f(\xi_i)\right)^6\right]\\ &\le \frac{e^{-3c^2}}{p^6}\left(p\mathbb{E}\left[f(\xi_i)^6\right] + 15p^2\mathbb{E}\left[f(\xi_i)^4\right]\mathbb{E}\left[f(\xi_i)^2\right] + 15p^3\mathbb{E}\left[f(\xi_i)^2\right]^3\right) < \frac{121e^{3c^2}}{p^3},\end{aligned}$$

where the last line holds for all $p \ge p_0(c)$. For any $i\neq l$, $(X_i^T X_l, X_l) \overset{L}{=} (\|X_l\|\xi_i, X_l)$ where $\xi_i\sim\mathcal{N}(0,1)$ is independent of X_l. Hence

$$\mathbb{P}\left[\frac{1}{p-1}\sum_{\substack{i=1\\ i\neq l}}^{p}\exp\left(\frac{16\beta}{\sqrt{n}}|X_i^T X_l|\right) > 3e^{\frac{128\beta^2\|X_l\|^2}{n}}\middle| X_l\right] < \frac{121e^{\frac{768\beta^2\|X_l\|^2}{n}}}{p^3}$$

for all $p \geq p_0\left(\beta, \frac{\|X_l\|}{\sqrt{n}}\right)$. Then

$$\mathbb{P}\left[\frac{1}{p-1}\sum_{\substack{i=1\\ i\neq l}}^{p}\exp\left(\frac{16\beta}{\sqrt{n}}|X_i^T X_l|\right) > 3e^{256\beta^2}\middle| \|X_l\| \leq \sqrt{2n}\right] < \frac{121e^{1536\beta^2}}{p^3}$$

for all $p \geq p_0(\beta)$, so using the chi-squared tail bound $\mathbb{P}[\|X_l\|^2 > 2n] \leq e^{-\frac{n}{8}}$,

$$\mathbb{P}\left[\frac{1}{p-1}\sum_{\substack{i=1\\ i\neq l}}^{p}\exp\left(\frac{16\beta}{\sqrt{n}}|X_i^T X_l|\right) > 3e^{256\beta^2}\right] \leq \frac{121e^{1536\beta^2}}{p^3} + e^{-\frac{n}{8}}.$$

The same argument holds for the analogous sums with ${X_i'}^T X_l$, $X_i^T X_l'$, and ${X_i'}^T X_l'$ in place of $X_i^T X_l$, and the result follows by a union bound. □

Lemma 6.5. *Let $y, z \in \mathbb{R}^p$ satisfy $\|y\| \leq 1$ and $\|z\| \leq 1$. Under the setup of Proposition 6.2, let $F(X) = z^T K_{n,p}(X) y$. Then $\mathbb{E}\left[e^{t(F(X)-F(X'))}\mathbb{1}\{(X, X') \in \mathcal{G}(\beta)\}\right] \leq 2\exp\left(\frac{C_{\beta,\gamma}\|y\|_\infty t^2}{\sqrt{n}}\right)$ for all $n \geq N_{\beta,\gamma}$, for constants $N_{\beta,\gamma}, C_{\beta,\gamma} > 0$ depending only on β and γ.*

Proof. Consider F as a function from $\mathbb{R}^{pn}$ to $\mathbb{R}$. Then

$$\begin{aligned}
\nabla_{X_l} F(X) &= \nabla_{X_l}\left(\sum_{i=1}^{p}\sum_{\substack{i'=1\\ i'\neq i}}^{p}\frac{1}{\sqrt{n}}k\left(\frac{X_i^T X_{i'}}{\sqrt{n}}\right) z_i y_{i'}\right)\\
&= \sum_{\substack{i=1\\ i\neq l}}^{p}\frac{1}{n}k'\left(\frac{X_i^T X_l}{\sqrt{n}}\right)(z_i y_l + y_i z_l) X_i^T\\
&= \frac{y_l}{n} v_z^T X + \frac{z_l}{n} v_y^T X,
\end{aligned}$$

for $v_y, v_z \in \mathbb{R}^p$ with $(v_y)_i = k'\left(\frac{X_i^T X_l}{\sqrt{n}}\right) y_i \mathbb{1}\{i \neq l\}$ and $(v_z)_i = k'\left(\frac{X_i^T X_l}{\sqrt{n}}\right) z_i \mathbb{1}\{i \neq l\}$. Then

$$\begin{aligned}
\|\nabla F(X)\|^2 &= \sum_{l=1}^{p}\|\nabla_{X_l} F(X)\|^2\\
&= \sum_{l=1}^{p}\frac{2y_l^2}{n^2}\|X\|^2\|v_z\|^2 + \frac{2z_l^2}{n^2}\|X\|^2\|v_y\|^2\\
&= \frac{4\|X\|^2}{n^2}\sum_{i=1}^{p}\sum_{\substack{l=1\\ l\neq i}}^{p}k'\left(\frac{X_i^T X_l}{\sqrt{n}}\right)^2 z_i^2 y_l^2\\
&\leq \frac{4\|X\|^2}{n^2}\max_{i=1}^{p}\sum_{\substack{l=1\\ l\neq i}}^{p}k'\left(\frac{X_i^T X_l}{\sqrt{n}}\right)^2 y_l^2\\
&\leq \frac{4\|X\|^2}{n^2}\max_{i=1}^{p}\left(\sum_{\substack{l=1\\ l\neq i}}^{p}k'\left(\frac{X_i^T X_l}{\sqrt{n}}\right)^4\right)^{1/2}\left(\sum_{\substack{l=1\\ l\neq i}}^{p} y_l^4\right)^{1/2}
\end{aligned}$$

$$\leq \frac{4\|X\|^2\|y\|_\infty}{n^2} \max_{i=1}^{p} \left(\sum_{\substack{l=1 \\ l\neq i}}^{p} k'\left(\frac{X_i^T X_l}{\sqrt{n}}\right)^4 \right)^{1/2} \left(\sum_{\substack{l=1 \\ l\neq i}}^{p} y_l^2 \right)^{1/2}$$

$$\leq \frac{4\|X\|^2\|y\|_\infty}{n^2} \max_{i=1}^{p} \left(\sum_{\substack{l=1 \\ l\neq i}}^{p} k'\left(\frac{X_i^T X_l}{\sqrt{n}}\right)^4 \right)^{1/2}.$$

For $\theta \in \left[0, \frac{\pi}{2}\right]$, let $X_\theta = X' \cos\theta + X \sin\theta$. Then

$$\|X_\theta\|^2 \leq (\|X'\cos\theta\| + \|X\sin\theta\|)^2 \leq 2\|X'\|^2(\cos\theta)^2 + 2\|X\|^2(\sin\theta)^2 \leq 2\max(\|X\|^2, \|X'\|^2),$$

so

$$\begin{aligned}
\sum_{\substack{l=1 \\ l\neq i}}^{p} k'\left(\frac{(X_\theta)_i^T (X_\theta)_l}{\sqrt{n}}\right)^4 &\leq \sum_{\substack{l=1 \\ l\neq i}}^{p} \exp\left(\frac{4\beta|(X_\theta)_i^T (X_\theta)_l|}{\sqrt{n}}\right) \\
&= \sum_{\substack{l=1 \\ l\neq i}}^{p} \exp\left(\frac{4\beta|(X_i'\cos\theta + X_i\sin\theta)^T (X_l'\cos\theta + X_l\sin\theta)|}{\sqrt{n}}\right) \\
&\leq \sum_{\substack{l=1 \\ l\neq i}}^{p} \exp\left(\frac{4\beta(|X_i^T X_l| + |X_i'^T X_l| + |X_i^T X_l'| + |X_i'^T X_l'|)}{\sqrt{n}}\right) \\
&\leq \left(\sum_{\substack{l=1 \\ l\neq i}}^{p} \exp\left(\frac{16\beta|X_i^T X_l|}{\sqrt{n}}\right) \right)^{1/4} \left(\sum_{\substack{l=1 \\ l\neq i}}^{p} \exp\left(\frac{16\beta|X_i'^T X_l|}{\sqrt{n}}\right) \right)^{1/4} \\
&\qquad \left(\sum_{\substack{l=1 \\ l\neq i}}^{p} \exp\left(\frac{16\beta|X_i^T X_l'|}{\sqrt{n}}\right) \right)^{1/4} \left(\sum_{\substack{l=1 \\ l\neq i}}^{p} \exp\left(\frac{16\beta|X_i'^T X_l'|}{\sqrt{n}}\right) \right)^{1/4},
\end{aligned}$$

where the last line follows from Hölder's inequality. Hence for any $(X, X') \in \mathcal{G}(\beta)$,

$$\|\nabla F(X_\theta)\|^2 \leq \frac{C_{\beta,\gamma}\|y\|_\infty}{n^{1/2}}$$

all $n \geq N_{\beta,\gamma}$ and some constants $N_{\beta,\gamma}, C_{\beta,\gamma} > 0$ depending on β and γ. Then

$$\begin{aligned}
&\mathbb{E}\left[e^{t(F(X)-F(X'))} \mathbb{1}\{(X, X') \in \mathcal{G}(\beta)\}\right] \\
&= \mathbb{E}\left[\exp\left(\frac{2}{\pi}\int_0^{\frac{\pi}{2}} \frac{\pi t}{2}\frac{d}{d\theta}F(X_\theta)d\theta\right) \mathbb{1}\{(X, X') \in \mathcal{G}(\beta)\}\right] \\
&\leq \mathbb{E}\left[\frac{2}{\pi}\int_0^{\frac{\pi}{2}} \exp\left(\frac{\pi t}{2}\frac{d}{d\theta}F(X_\theta)\right) d\theta\, \mathbb{1}\{(X, X') \in \mathcal{G}(\beta)\}\right] \\
&= \frac{2}{\pi}\int_0^{\frac{\pi}{2}} \mathbb{E}\left[\exp\left(\frac{\pi t}{2}\nabla F(X_\theta)^T \tilde{X}_\theta\right) \mathbb{1}\{(X, X') \in \mathcal{G}(\beta)\}\right] d\theta,
\end{aligned}$$

where the third line follows from Jensen's inequality, $\tilde{X}_\theta = -X' \sin\theta + X\cos\theta$ in the fourth line, and the inner-product $\nabla F(X_\theta)^T \tilde{X}_\theta$ represents the vector inner-product in $\mathbb{R}^{pn}$. Noting that X_θ and $\tilde{X}_\theta$ are independent and both equal in law to X, we may first condition on X_θ and use the Cauchy-Schwarz inequality and $\mathbb{E}\left[e^{c|(\tilde{X}_\theta)_{ij}|}\right] \leq \mathbb{E}\left[e^{c(\tilde{X}_\theta)_{ij}}\right] + \mathbb{E}\left[e^{-c(\tilde{X}_\theta)_{ij}}\right] \leq 2e^{\frac{c^2}{2}}$ to obtain

$$
\begin{aligned}
&\mathbb{E}\left[e^{t(F(X)-F(X'))}\mathbb{1}\{(X,X') \in \mathcal{G}(\beta)\}\right] \\
&= \frac{2}{\pi}\int_0^{\frac{\pi}{2}} \mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{\pi t}{2}\nabla F(X_\theta)^T \tilde{X}_\theta\right)\mathbb{1}\{(X,X') \in \mathcal{G}(\beta)\}\middle| X_\theta\right]\right] d\theta \\
&\leq \frac{2}{\pi}\int_0^{\frac{\pi}{2}} \mathbb{E}\left[\mathbb{E}\left[\exp\left(\pi t \nabla F(X_\theta)^T \tilde{X}_\theta\right)\middle| X_\theta\right]^{\frac{1}{2}} \mathbb{E}\left[\mathbb{1}\{(X,X') \in \mathcal{G}(\beta)\}\middle| X_\theta\right]^{\frac{1}{2}}\right] d\theta \\
&\leq \frac{4}{\pi}\int_0^{\frac{\pi}{2}} \mathbb{E}\left[\exp\left(\frac{\pi^2 t^2 \|\nabla F(X_\theta)\|^2}{4}\right)\mathbb{E}\left[\mathbb{1}\{(X,X') \in \mathcal{G}(\beta)\}\middle| X_\theta\right]^{\frac{1}{2}}\right] d\theta \\
&= \frac{4}{\pi}\int_0^{\frac{\pi}{2}} \mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{\pi^2 t^2 \|\nabla F(X_\theta)\|^2}{2}\right)\mathbb{1}\{(X,X') \in \mathcal{G}(\beta)\}\middle| X_\theta\right]^{\frac{1}{2}}\right] d\theta \\
&\leq \frac{4}{\pi}\int_0^{\frac{\pi}{2}} \mathbb{E}\left[\exp\left(\frac{\pi^2 t^2 \|\nabla F(X_\theta)\|^2}{2}\right)\mathbb{1}\{(X,X') \in \mathcal{G}(\beta)\}\right]^{\frac{1}{2}} d\theta \\
&\leq 2\exp\left(\frac{\pi^2 C_{\beta,\gamma}\|y\|_\infty t^2}{4n^{1/2}}\right).
\end{aligned}
$$

□

Definition 6.6. *For $m = \lceil \log_2 p \rceil$, let $D_2^p = \left\{y \in \mathbb{R}^p : \|y\| \leq 1, \forall i, y_i^2 \in \{0, 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots, \frac{1}{2^{m+3}}\}\right\}$. For each $l = 0, 1, \ldots, m+3$, let $\pi_l : D_2^p \to D_2^p$ be defined by $(\pi_l(y))_i = y_i \mathbb{1}\{y_i^2 \geq 2^{-l}\}$, and let $\pi_{l\backslash l-1} : D_2^p \to D_2^p$ be defined by $(\pi_{l\backslash l-1}(y))_i = y_i \mathbb{1}\{y_i^2 = 2^{-l}\}$.*

Lemma 6.7. *For any $x \in \mathbb{R}^p$ with $\|x\| \leq 1$, there exists $y \in D_2^p$ such that $\|x - y\| < \frac{9}{20}$, and hence $\|M\| \leq 10 \sup_{y \in D_2^p} y^T M y$ for any symmetric $M \in \mathbb{R}^{p \times p}$.*

Proof. If $x = 0$ or $x_i^2 = 1$ for some i, then we may take $y = x$. Otherwise, for each $i = 1, \ldots, p$, if $2^{-l} \leq x_i^2 < 2^{-l+1}$ for some $l \in \{1, \ldots, m+3\}$, let $y_i = 2^{-\frac{l}{2}} \operatorname{sign}(x_i)$, and if $x_i^2 < 2^{-m-3}$, let $y_i = 0$. Then $\|y\| \leq \|x\| \leq 1$ and $y \in D_2^p$, and

$$
\begin{aligned}
\|x - y\|^2 &= \sum_{i: x_i^2 \geq 2^{-m-3}} (x_i - y_i)^2 + \sum_{i: x_i^2 < 2^{-m-3}} x_i^2 \\
&\leq \sum_{i: x_i^2 \geq 2^{-m-3}} \left(1 - \frac{1}{\sqrt{2}}\right)^2 x_i^2 + \sum_{i: x_i^2 < 2^{-m-3}} x_i^2 \\
&\leq \left(1 - \frac{1}{\sqrt{2}}\right)^2 + \left(1 - \left(1 - \frac{1}{\sqrt{2}}\right)^2\right) \sum_{i: x_i^2 < 2^{-m-3}} x_i^2 \\
&\leq \left(1 - \frac{1}{\sqrt{2}}\right)^2 + \frac{1}{8}\left(1 - \left(1 - \frac{1}{\sqrt{2}}\right)^2\right) \\
&< \left(\frac{9}{20}\right)^2
\end{aligned}
$$

establishing the first claim. The second claim then follows from Lemma 5.4 in [31]. □

Lemma 6.8. *For some constant $C > 0$, $m = \lceil \log_2 p \rceil$, and all $l \in \{0, 1, \dots, m+3\}$, $\log |\{\pi_l(y) : y \in D_2^p\}| \le C(m+4-l)2^l$.*

Proof. For any $l \in \{0, 1, \dots, m\}$,

$$|\{\pi_{l\backslash l-1}(y) : y \in D_2^p\}| \le \sum_{k=0}^{2^l} \binom{p}{k} 2^k,$$

as there are at most 2^l non-zero entries of $\pi_{l\backslash l-1}(y)$, and for each non-zero entry there are two choices of sign. Using $\binom{p}{k} \le \left(\frac{ep}{k}\right)^k$, and noting that $k \mapsto (2ep)^k k^{-k}$ is monotonically increasing over $k \in [0, 2p]$ and that $2^l \le 2p$ for $l \le m$, this implies

$$\log |\{\pi_{l\backslash l-1}(y) : y \in D_2^p\}| \le \log \left(1 + 2^l \left(\frac{2ep}{2^l}\right)^{2^l}\right) \le \log \left(1 + 2^l \left(2e2^{m-l}\right)^{2^l}\right) \le C(m-l+1)2^l$$

for a constant $C > 0$. For $l \in \{m+1, m+2, m+3\}$, we use the bound $|\{\pi_{l\backslash l-1}(y) : y \in D_2^p\}| \le 3^p$, as each coordinate of $\pi_{l\backslash l-1}(y)$ takes one of three values. Then

$$\log |\{\pi_{l\backslash l-1}(y) : y \in D_2^p\}| \le C2^m \le C(m+4-l)2^l$$

for a constant $C > 0$. Then

$$\begin{aligned}\log |\pi_l(y)| &\le \sum_{j=0}^{l} \log |\pi_{j\backslash j-1}(y)| \le C \sum_{j=0}^{l} (m+4-j)2^j = C(m+4-l) \sum_{j=0}^{l} 2^j + C \sum_{k=0}^{l-1} \sum_{j=0}^{k} 2^j \\ &\le 2C(m+4-l)2^l + 2C2^l \le C'(m+4-l)2^l\end{aligned}$$

for a constant $C' > 0$. □

Lemma 6.9. *Under the setup of Proposition 6.2, let $m = \lceil \log_2 p \rceil$. Then there are constants $C, C_{\beta,\gamma}, N_{\beta,\gamma} > 0$ such that for any $l \in \{0, 1, \dots, m+3\}$, any $t > 0$, and any $n \ge N_{\beta,\gamma}$,*

$$\mathbb{P}\left[\sup_{y \in D_2^p} \pi_l(y)^T K_{n,p}(X) \pi_{l\backslash l-1}(y) > t \text{ and } (X, X') \in \mathcal{G}(\beta)\right] \le 2e^{C(m+4-l)2^l - \frac{t^2\sqrt{2^l n}}{4C_{\beta,\gamma}}},$$

and for any $l \in \{1, \dots, m+3\}$, any $t > 0$, and any $n \ge N_{\beta,\gamma}$,

$$\mathbb{P}\left[\sup_{y \in D_2^p} \pi_{l-1}(y)^T K_{n,p}(X) \pi_{l\backslash l-1}(y) > t \text{ and } (X, X') \in \mathcal{G}(\beta)\right] \le 2e^{C(m+4-l)2^l - \frac{t^2\sqrt{2^l n}}{4C_{\beta,\gamma}}}.$$

Proof. Letting $m = \lceil \log_2 p \rceil$ and $j = l-1$ or l,

$$\begin{aligned}&\mathbb{P}\left[\sup_{y \in D_2^p} \pi_j(y) K_{n,p}(X) \pi_{l\backslash l-1}(y) > t \text{ and } (X, X') \in \mathcal{G}(\beta)\right] \\ &= \mathbb{P}\left[\sup_{y \in \{\pi_l(x) : x \in D_2^p\}} \pi_j(y) K_{n,p}(X) \pi_{l\backslash l-1}(y) > t \text{ and } (X, X') \in \mathcal{G}(\beta)\right] \\ &\le e^{C(m+4-l)2^l} \sup_{y \in \{\pi_l(x) : x \in D_2^p\}} \mathbb{P}\left[\pi_j(y) K_{n,p}(X) \pi_{l\backslash l-1}(y) > t \text{ and } (X, X') \in \mathcal{G}(\beta)\right] \\ &\le e^{C(m+4-l)2^l} e^{-\lambda t} \sup_{y \in \{\pi_l(x) : x \in D_2^p\}} \mathbb{E}\left[e^{\lambda \pi_j(y) K_{n,p}(X) \pi_{l\backslash l-1}(y)} \mathbb{1}\{(X, X') \in \mathcal{G}(\beta)\}\right],\end{aligned}$$

where the second line follows from $\pi_{l\backslash l-1}(y) = \pi_{l\backslash l-1}(\pi_l(y))$ and $\pi_{l-1}(y) = \pi_{l-1}(\pi_l(y))$, the third line follows from a union bound and Lemma 6.8 for some constant $C > 0$, and the fourth line

follows from Markov's inequality for any $\lambda > 0$. Let Λ be the set of all diagonal matrices in $\mathbb{R}^{p\times p}$ with all diagonal entries in $\{-1, 1\}$. Note that $(X, X') \in \mathcal{G}(\beta)$ if and only if $(X, DX') \in \mathcal{G}(\beta)$ for all $D \in \Lambda$. Then, conditional on X and $(X, X') \in \mathcal{G}(\beta)$, X' is equal in law to DX' for D uniformly distributed over Λ, and

$$\begin{aligned}
&\mathbb{E}[K_{n,p}(X')|X, (X, X') \in \mathcal{G}(\beta)] \\
&= \mathbb{E}[K_{n,p}(DX')|X, (X, X') \in \mathcal{G}(\beta)] \\
&= \mathbb{E}[\mathbb{E}[K_{n,p}(DX')|X', X, (X, X') \in \mathcal{G}(\beta)]|X, (X, X') \in \mathcal{G}(\beta)] \\
&= 0,
\end{aligned}$$

where the last line follows from $\mathbb{E}[K_{n,p}(DX')] = 0$ for any fixed $X' \in \mathbb{R}^{p\times n}$ as the kernel function k is odd. Hence by Jensen's inequality, for any $y \in D_2^p$,

$$\mathbb{E}\left[e^{-\lambda\pi_j(y)K_{n,p}(X')\pi_{l\backslash l-1}(y)}\middle|X, (X, X') \in \mathcal{G}(\beta)\right] \geq 1,$$

and so

$$\begin{aligned}
&\mathbb{E}\left[e^{\lambda\pi_j(y)K_{n,p}(X)\pi_{l\backslash l-1}(y)}\mathbb{1}\{(X, X') \in \mathcal{G}(\beta)\}\right] \\
&= \mathbb{E}\left[e^{\lambda\pi_j(y)K_{n,p}(X)\pi_{l\backslash l-1}(y)}\middle|(X, X') \in \mathcal{G}(\beta)\right]\mathbb{P}[(X, X') \in \mathcal{G}(\beta)] \\
&\leq \mathbb{E}\left[e^{\lambda\pi_j(y)K_{n,p}(X)\pi_{l\backslash l-1}(y)}\mathbb{E}\left[e^{-\lambda yK_{n,p}(X')\pi_{l\backslash l-1}(y)}\middle|X, (X, X') \in \mathcal{G}(\beta)\right]\middle|(X, X') \in \mathcal{G}(\beta)\right]\mathbb{P}[(X, X') \in \mathcal{G}(\beta)] \\
&= \mathbb{E}\left[e^{\lambda\pi_j(y)(K_{n,p}(X)-K_{n,p}(X'))\pi_{l\backslash l-1}(y)}\middle|(X, X') \in \mathcal{G}(\beta)\right]\mathbb{P}[(X, X') \in \mathcal{G}(\beta)] \\
&= \mathbb{E}\left[e^{\lambda\pi_j(y)(K_{n,p}(X)-K_{n,p}(X'))\pi_{l\backslash l-1}(y)}\mathbb{1}\{(X, X') \in \mathcal{G}(\beta)\}\right].
\end{aligned}$$

Noting that $\|\pi_{l\backslash l-1}(y)\|_\infty \leq 2^{-\frac{l}{2}}$, Lemma 6.5 implies this last quantity is at most $2\exp\left(\frac{C_{\beta,\gamma}\lambda^2}{\sqrt{2^l n}}\right)$ for all $y \in D_2^p$ and $n \geq N_{\beta,\gamma}$. Setting $\lambda = \frac{t\sqrt{2^l n}}{2C_{\beta,\gamma}}$ gives the desired result. □

Proof of Proposition 6.2. Let $m = \lceil \log_2 p \rceil$, let D_2^p, π_l, and $\pi_{l\backslash l-1}$ be as in Definition 6.6, and let $C, C_{\beta,\gamma} > 0$ be the constants in Lemma 6.9. We may assume without loss of generality $C \geq 3$. For each $l = 0, \ldots, m+3$, let

$$t_l = \frac{\sqrt{8CC_{\beta,\gamma}(m+4-l)}2^{\frac{l}{4}}}{n^{\frac{1}{4}}}.$$

Let X' be an independent copy of X. Then by Lemma 6.9, for all $n \geq N_{\beta,\gamma}$ and each $l = 0, \ldots, m+3$,

$$\mathbb{P}\left[\sup_{y\in D_2^p}\pi_l(y)^T K_{n,p}(X)\pi_{l\backslash l-1}(y) > t_l \text{ and } (X, X') \in \mathcal{G}(\beta)\right] \leq 2e^{-C(m+4-l)2^l},$$

and for each $l = 1, \ldots, m+3$,

$$\mathbb{P}\left[\sup_{y\in D_2^p}\pi_{l-1}(y)^T K_{n,p}(X)\pi_{l\backslash l-1}(y) > t_l \text{ and } (X, X') \in \mathcal{G}(\beta)\right] \leq 2e^{-C(m+4-l)2^l}.$$

Note

$$2\sum_{l=0}^{m+3} t_l \leq \frac{2\sqrt{8CC_{\beta,\gamma}}}{n^{\frac{1}{4}}}\sum_{l=0}^{m+3}(m+4-l)2^{\frac{l}{4}} = \frac{2\sqrt{8CC_{\beta,\gamma}}}{n^{\frac{1}{4}}}\sum_{l=0}^{m+3}\sum_{j=0}^{l}2^{\frac{j}{4}} \leq \frac{C'_{\beta,\gamma}2^{\frac{m}{4}}}{n^{\frac{1}{4}}} \leq C''_{\beta,\gamma}$$

for some constants $C'_{\beta,\gamma}, C''_{\beta,\gamma} > 0$, while

$$\sup_{y\in D_2^p} y^T K_{n,p}(X)y = \sup_{y\in D_2^p}\left(\sum_{l=0}^{m+3}\pi_l(y)^T K_{n,p}(X)\pi_{l\backslash l-1}(y) + \sum_{l=1}^{m+3}\pi_{l\backslash l-1}(y)^T K_{n,p}(X)\pi_{l-1}(y)\right)$$

$$\leq \sum_{l=0}^{m+3} \sup_{y \in D_2^p} \pi_l(y)^T K_{n,p}(X) \pi_{l\backslash l-1}(y) + \sum_{l=1}^{m+3} \sup_{y \in D_2^p} \pi_{l-1}(y)^T K_{n,p}(X) \pi_{l\backslash l-1}(y).$$

Then

$$\begin{aligned}
&\mathbb{P}\left[\sup_{y \in D_2^p} y^T K_{n,p}(X) y > C'''_{\beta,\gamma} \text{ and } (X, X') \in \mathcal{G}(\beta)\right] \\
&\leq \sum_{l=0}^{m+3} \mathbb{P}\left[\sup_{y \in D_2^p} \pi_l(y)^T K_{n,p}(X) \pi_{l\backslash l-1}(y) > t_l \text{ and } (X, X') \in \mathcal{G}(\beta)\right] \\
&\qquad + \sum_{l=1}^{m+3} \mathbb{P}\left[\sup_{y \in D_2^p} \pi_{l-1}(y)^T K_{n,p}(X) \pi_{l\backslash l-1}(y) > t_l \text{ and } (X, X') \in \mathcal{G}(\beta)\right] \\
&\leq 4 \sum_{l=0}^{m+3} e^{-C(m+4-l)2^l} < 4(m+4)e^{-3m} < \frac{C' \log p}{p^3}
\end{aligned}$$

for some constant $C' > 0$ and all $n \geq N_{\beta,\gamma}$. By Lemma 6.4, this implies

$$\mathbb{P}\left[\sup_{y \in D_2^p} y^T K_{n,p}(X) y > C''_{\beta,\gamma}\right] < \frac{C' \log p}{p^3} + \mathbb{P}[(X, X)' \notin \mathcal{G}(\beta)] \leq \frac{C'''_{\beta,\gamma}}{n^2}$$

for all $n \geq N'_{\beta,\gamma}$ and some constants $C'''_{\beta,\gamma}, N'_{\beta,\gamma} > 0$. The desired result then follows from Lemma 6.7. □

Appendix A. Combinatorial results

This appendix contains the proofs of a number of combinatorial lemmas used in Section 4, as well as the proof of Proposition 4.15 and the explicit construction of the map φ in that proposition. We restate any lemmas previously stated in Section 4 using their original numbering.

Lemma 4.11. *Suppose a simple-labeling of an l-graph has $\tilde{k}$ n-vertices with non-empty label and $\tilde{m}$ total distinct p-labels and distinct non-empty n-labels. Then $\tilde{m} \leq \frac{l+\tilde{k}}{2} + 1$.*

Proof. Let $I = \{1, \ldots, p\}$ and $J = \{1, \ldots, n\}$, and consider an undirected graph G on the vertex set $I \sqcup J$ (the disjoint union of I and J with $n+p$ elements, treating elements of I and the elements of J as distinct). Let G have an edge between $i, i' \in I$ if there are two consecutive p-vertices of the l-graph having p-labels i and i' in the simple-labeling and such that the n-vertex between them has empty label, and let G have an edge between $i \in I$ and $j \in J$ if there are two consecutive vertices of the l-graph such that the p-vertex has label i and the n-vertex has label j. The number of vertices of G incident to at least one edge is $\tilde{m}$, and G must be connected, so it has at least $\tilde{m} - 1$ edges. Note that an edge in G between $i, i' \in I$ corresponds to at least two consecutive pairs of p-vertices in the l-graph such that the n-vertex between them has empty label, by condition (3) of Definition 4.10, so the number of such edges is at most $\frac{l-\tilde{k}}{2}$. Similarly, an edge in G between $i \in I$ and $j \in J$ corresponds to at least two pairs of consecutive n and p-vertices of the l-graph such that the n-vertex has non-empty label, by condition (2) of 4.10. Hence, the number of such edges is at most $\frac{2\tilde{k}}{2}$. Then $\tilde{m} - 1 \leq \frac{l+\tilde{k}}{2}$. □

Lemma A.1. *In any multi-labeling of an l-graph, each distinct n-label that appears must appear on at least two n-vertices.*

Proof. Suppose that an n-label j appears only once. The two p-vertices preceding and following that n-vertex must have distinct labels, say i_1 and i_2, by condition (1) of Definition 4.3. Then

exactly one edge in the l-graph has p-vertex endpoint labeled i_1 and n-vertex endpoint having label j (and similarly for i_2 and j), contradicting condition (3) of Definition 4.3. □

Lemma A.2. *Suppose $l = 2$ or $l = 3$. Then for any multi-labeling of the l-graph, all l p-labels are distinct, and all l n-vertices have the same tuple of n-labels, up to reordering.*

Proof. That all l p-labels are distinct is a consequence of condition (1) of Definition 4.3. Then by conditions (2) and (3) of Definition 4.3, the n-vertices immediately preceding and following each p-vertex must have the same tuple of n-labels, up to reordering, and hence all l n-vertices has the same tuple of n-labels. □

Lemma A.3. *In a multi-labeling of an l-graph with $l \geq 4$, suppose a p-vertex V is such that its p-label appears on no other p-vertices. Let the n-vertex preceding V be U, the p-vertex preceding U be T, the n-vertex following V be W, and the p-vertex following W be X.*

1. *If T and X have different p-labels, then the graph obtained by deleting V and W and connecting U to X is an $(l-1)$-graph with valid multi-labeling.*
2. *If T and X have the same p-label, then the graph obtained by deleting U, V, W, and X and connecting U to the n-vertex after X is an $(l-2)$-graph with valid multi-labeling.*

Proof. First consider case (1). As T and X have distinct p-labels, it remains true that no two consecutive p-vertices in the $(l-1)$-graph have the same p-label, so condition (1) of Definition 4.3 holds. Condition (2) of Definition 4.3 clearly still holds as well. If V has p-label i and W has n-labels $(j_1, \ldots, j_d)$, then U has n-labels $(j_1, \ldots, j_d)$ as well, up to reordering, by conditions (2) and (3) of Definition 4.3 for the original l-graph and the fact that V is the only p-vertex with label i. Then in the $(l-1)$-graph obtained by deleting V and W, the number of edges with p-vertex endpoint labeled i and n-vertex endpoint having label j_s for any $s = 1, \ldots, d$ is zero, and the number of edges with p-vertex endpoint labeled i' and n-vertex endpoint having label j' is the same as in the original l-graph for all other pairs (i', j'). Thus condition (3) of Definition 4.3 still holds as well, so the $(l-1)$-graph still has a valid multi-labeling.

Now consider case (2). X and the p-vertex after X must have different p-labels in the original l-graph, by condition (1) of Definition 4.3. As T and X have the same p-label, this implies T and the p-vertex after X must have different p-labels, so condition (1) of Definition 4.3 still holds in the $(l-2)$-graph. Condition (2) of Definition 4.3 clearly still holds in the $(l-2)$-graph as well. Suppose V has p-label i_1, T and X have p-label i_2, and W has n-labels $(j_1, \ldots, j_d)$. Then by conditions (2) and (3) of Definition 4.3 for the original l-graph and the fact that V is the only p-vertex with label i_1, U must also have n-labels $(j_1, \ldots, j_d)$, up to reordering. Then in the $(l-2)$-graph obtained by deleting U, V, W, and X, the number of edges with p-vertex endpoint labeled i_1 and n-vertex endpoint having label j_s for any $s = 1, \ldots, d$ is zero, the number of edges with p-vertex endpoint labeled i_2 and n-vertex endpoint having label j_s for any $s = 1, \ldots, d$ is two less than in the original l-graph, and the number of edges with p-vertex endpoint labeled i' and n-vertex endpoint having label j' is the same as in the original l-graph for all other pairs (i', j'). Hence condition (3) of Definition 4.3 still holds as well, so the $(l-2)$-graph still has a valid multi-labeling. □

Lemma 4.4. *Suppose a multi-labeling of an l-graph has $d_1, \ldots, d_l$ n-labels on the first through l^{th} n-vertices, respectively, and suppose that it has m total distinct p-labels and n-labels. Then $m \leq \frac{l+\sum_{s=1}^{l} d_s}{2} + 1$.*

Proof. We induct on l. For $l = 2$, a multi-labeling must have $d_1 = d_2$ and $m = d_1 + 2$, and for $l = 3$, a multi-labeling must have $d_1 = d_2 = d_3$ and $m = d_1 + 3$, by Lemma A.2. The result is then easily verified in these two cases.

Suppose by induction that the result holds for $l-2$ and $l-1$, and consider a multi-labeling of an l-graph with $d_1, \ldots, d_l$ and m as specified, and $l \geq 4$. If each distinct p-label which appears in

the labeling appears at least twice, then there are at most $\frac{l}{2}$ distinct p-labels. Lemma A.1 implies there are at most $\frac{\sum_{s=1}^{l} d_s}{2}$ distinct n-labels, so $m \le \frac{l+\sum_{s=1}^{l} d_s}{2}$, establishing the result.

Thus, suppose that some p-vertex V has a label that appears exactly once. Let the n-vertex preceding V be U, the p-vertex preceding U be T, the n-vertex following V be W, and the p-vertex following W be X. If T and X have different p-labels, follow procedure (1) in Lemma A.3 to obtain a multi-labeling of an $(l-1)$-graph. This multi-labeling now has $m-1$ total distinct p-labels and n-labels, and so the induction hypothesis implies $m-1 \le \frac{l-1+\sum_{s=1}^{l} d_s - d}{2} + 1$ where d is the number of n-labels of the deleted n-vertex W. Hence $m \le \frac{l+\sum_{s=1}^{l} d_s}{2} - \frac{d+1}{2} + 2 \le \frac{l+\sum_{s=1}^{l} d_s}{2} + 1$.

If T and X have the same p-label, follow procedure (2) of Lemma A.3 to obtain a multi-labeling of an $(l-2)$-graph. This multi-labeling has at least $m-d-1$ and at most $m-1$ total distinct p-labels and n-labels, where d is the number of n-labels of the deleted n-vertex W. The induction hypothesis implies $m-d-1 \le \frac{l-2+\sum_{s=1}^{l} d_s - 2d}{2} + 1$, so $m \le \frac{l+\sum_{s=1}^{l} d_s}{2} + 1$. This completes the induction in both cases, establishing the desired result. □

Lemma A.4. *Suppose a multi-labeling of an l-graph has excess Δ, $d_1, \dots, d_l$ n-labels on the first through l^{th} n-vertices, respectively, and at most $\frac{l}{2}$ distinct p-labels. Then at most $6\Delta - 6$ of the $\sum_{s=1}^{l} d_s$ total n-labels are such that that distinct n-label appears three or more times in the labeling.*

Proof. Observe that if m total distinct p-labels and n-labels appear in the labeling, and at most $\frac{l}{2}$ of these are p-labels, then the labeling has at least $m - \frac{l}{2}$ distinct n-labels. If c is the number of distinct n-labels that appear exactly twice, then, as each distinct n-label appears at least twice by Lemma A.1, a pigeonhole argument implies $2c + 3\left(m - \frac{l}{2} - c\right) \le \sum_{s=1}^{l} d_s$, so $c \ge 3m - \frac{3l}{2} - \sum_{s=1}^{l} d_s$. Then the n-labels which appear exactly twice account for at least $6m - 3l - 2\sum_{s=1}^{l} d_s$ of the $\sum_{s=1}^{l} d_s$ total n-labels, implying that at most $3l + 3\sum_{s=1}^{l} d_s - 6m = 6\Delta - 6$ total n-labels remain. □

Remark A.5. *Lemma A.4 implies that if an l-graph has at most $\frac{l}{2}$ distinct p-labels, then it has excess $\Delta \ge 1$.*

Lemma A.6. *Suppose a multi-labeling of an l-graph has excess Δ. For each p-label i and n-label j, let b_{ij} be the number of pairs of consecutive vertices in the l-graph such that the p-vertex of the pair is labeled i and the n-vertex of the pair has label j in its tuple. Then $\sum_{i,j:b_{ij}>2} b_{ij} \le 12\Delta$.*

Proof. We induct on l. For $l = 2$ or 3, we must have $b_{ij} = 2$ for all (i, j) by Lemma A.2, and $\Delta \ge 0$ by Lemma 4.4, so the result holds.

Suppose the result holds for $l-2$ and $l-1$, and consider a multi-labeling of an l-graph with $l \ge 4$. If each p-label that appears in the labeling appears at least twice, then there are at most $\frac{l}{2}$ distinct p-labels, so Lemma A.4 implies that at most 6Δ of the total n-labels are such that that distinct n-label appears at least three times. Note that for any distinct n-label j that appears only twice, $b_{ij} = 2$ or $b_{ij} = 0$ for all p-labels i, by conditions (1) and (3) of Definition 4.3. For any distinct n-label j that appears at least three times, $\sum_{i:b_{ij}>2} b_{ij}$ is exactly twice the total number of appearances of j as an n-label (as the p-vertices to the left and right of any n-vertex containing label j each contributes 1 to this sum). Then $\sum_{i,j:b_{ij}>2} b_{ij} \le 12\Delta$ by Lemma A.4.

Now suppose that some p-vertex V has a label i that appears exactly once in the labeling. Consider the $(l-1)$-graph or $(l-2)$-graph obtained by following either procedure (1) or procedure (2) of Lemma A.3. In the $(l-1)$-graph, $b_{ij} = 0$ for n-labels j that appear on the deleted n-vertices U and W, whereas $b_{ij} = 2$ for such n-labels j in the original l-graph, and the values $b_{i'j'}$ for all other pairs (i', j') are the same in the two graphs. Hence $\sum_{i,j:b_{ij}>2} b_{ij}$ is the same in the $(l-1)$-graph as in the l-graph. Then the induction hypothesis implies $\sum_{i,j:b_{ij}>2} b_{ij} \le 12\left(\frac{l-1+\sum_{s=1}^{l} d_s - d}{2} + 1 - (m-1)\right) \le 12\Delta$, where d is the number of n-labels on the deleted n-vertex W (or U) and m is the total number of distinct n and p-labels in the original l-graph.

In the case of the $(l-2)$-graph, suppose the deleted n-vertex W (or U) had d n-labels, of which d' appear on an n-vertex other than W and U. If j is a distinct n-label that does not appear on W or U, then clearly b_{ij} is the same in the $(l-2)$-graph and the original l-graph for all i. If j is one of the $d-d'$ distinct n-labels that appear only on W and U, then $b_{ij} = 0$ or 2 in both the $(l-2)$-graph and the original l-graph for all i. If j is one of the other d' distinct n-labels, then in deleting U, V, W, and X, we may have reduced b_{ij} by 2 for at most two distinct i-labels (corresponding to the i-labels of V and X). This implies $\sum_{i:b_{ij}>2} b_{ij}$ reduces by at most 8 for this j, with the maximal reduction occurring if $b_{ij} = 4$ for both of these distinct i-labels in the original l-graph. Then by the induction hypothesis, $\sum_{i,j:b_{ij}>2} b_{ij} - 8d' \leq 12\left(\frac{l-2+\sum_{s=1}^{l} d_s - 2d}{2} + 1 - (m-1-(d-d')\right)$, as the $(l-2)$-graph has $m-1-(d-d')$ total distinct n and p-labels. Then $\sum_{i,j:b_{ij}>2} b_{ij} \leq 12\left(\frac{l+\sum_{s=1}^{l} d_s}{2} + 1 - m - d'\right) + 8d' \leq 12\Delta$, so the result holds in this case as well, completing the induction. □

This concludes the proof of all combinatorial lemmas used in Section 4. The remainder of this appendix establishes Proposition 4.15 and provides the explicit construction of the map φ in that proposition.

Definition A.7. *In an l-graph with a multi-labeling, an n-vertex is* **single** *if it has only one n-label. It is a* **good single** *if it is single and if its n-label does not appear on any other n-vertex that is not single. Otherwise, it is a* **bad single**.

Definition A.8. *In an l-graph with a (p,n)-multi-labeling, a pair (V,V') of distinct (not necessarily consecutive) n-vertices is a* **good pair** *if the following conditions hold:*

1. *V and V' have the same tuple of n-labels, up to reordering,*
2. *V has at least two n-labels (as does V'),*
3. *Each distinct n-label appearing on V and V' does not appear on any other n-vertices besides V and V'.*

Remark A.9. *The p-labels of the two p-vertices preceding and following V must be distinct, by condition (1) of Definition 4.3, and similarly for the p-labels of the two p-vertices preceding and following V'. As each n-label for V and V' does not appear on any other vertices besides V and V', condition (3) of Definition 4.3 requires that, in fact, the two p-labels of the p-vertices preceding and following V are the same as those of the p-vertices preceding and following V' (but not necessarily in the same order).*

Definition A.10. *Suppose (V,V') is a good pair of n-vertices. Let the p-vertices preceding and following V be U and W, respectively, and let the p-vertices preceding and following V' be U' and W', respectively. Then the good pair (V,V') is* **proper** *if U has the same label as W' and U' has the same label as W, and it is* **improper** *if U has the same label as U' and W has the same label as W'.*

Definition A.11. *The* **label-simplifying map** *is the map from (p,n)-multi-labelings of an l-graph to $(p,n+1)$-simple-labelings of an l-graph, defined by the following procedure:*

1. *While there exists an improper good pair of n-vertices (V,V'), iterate the following: Let W be the p-vertex following V and W' be the p-vertex following V', and reverse the sequence of vertices starting at W and ending at W' along with their labels. (I.e., swap W with W', the n-vertex following W with the n-vertex preceding W', etc.)*
2. *For each n-vertex in a good pair, relabel it with the empty label.*
3. *For each n-vertex that is neither a good single nor part of a good pair, relabel it with the single label $n+1$.*

Remark A.12. *In the case where there are multiple improper good pairs in step (1) of this procedure, it will not be important for our later arguments in which order the pairs (V,V') are selected.*

For concreteness, we may always select $\{V, V'\}$ to be the improper good pair whose sorted n-label-tuple is smallest lexicographically, and we may take V to come before V' in the l-graph cycle.

Lemma A.13. *The following are true for the label-simplifying map in Definition A.11:*

(1) *Step (1) of the procedure in Definition A.11 always terminates in a valid (p, n)-multi-labeling with no improper good pairs.*
(2) *The image of any (p, n)-multi-labeling under the map is a valid $(p, n + 1)$-simple-labeling.*
(3) *If two multi-labelings are equivalent, then their image simple-labelings are also equivalent.*

Proof. Clearly each reversal in step (1) of the procedure in Definition A.11 preserves condition (2) of Definition 4.3 as well as the number of good pairs and n-labels of each good pair. As W and W' have the same p-label because (V, V') is improper, it also preserves conditions (1) and (3) of Definition 4.3, so the labeling of the vertices after each such reversal is still a valid (p, n)-multi-labeling. Each time an improper good pair (V, V') is identified and a reversal is performed, V and V' become consecutive n-vertices in the l-graph, and the pair (V, V') becomes a proper good pair. As V and V' are consecutive, they must remain consecutive under each subsequent reversal that is performed, so their properness is preserved. Hence the procedure must terminate after a number of iterations at most the total number of good pairs in the multi-labeling, and the final multi-labeling is such that all good pairs are proper. This establishes (1).

To prove (2), note that the image labeling under the label-simplifying map has either one n-label or the empty label for each n-vertex. Condition (1) of Definition 4.10 holds for the image labeling by condition (1) of Definition 4.3 for multi-labelings, as the p-labels are preserved under steps (2) and (3) of the procedure in Definition A.11. As all good pairs in the multi-labeling obtained after applying step (1) of the procedure are proper, and step (2) of the procedure maps their labels to the empty label, condition (3) of Definition 4.10 holds for the image simple-labeling. Finally, note that if j is an n-label appearing on good single vertices in the multi-labeling, then condition (2) of Definition 4.10 holds in the image labeling for this j and all p-labels i by condition (3) in Definition 4.3 for multi-labelings. For the new n-label $n + 1$ created under the label-simplifying map, note that for each distinct p-label i, there must be an even number of total edges in the l-graph cycle with p-endpoint labeled i. Of these, there must be an even number with n-endpoint j for any good single label j, by the above argument, and there must also be an even number with n-endpoint belonging to a good pair in the multi-labeling since these edges must come in pairs. Hence the number of remaining edges adjacent to any p-vertex with label i must also be even. These are precisely the edges with p-vertex endpoint labeled i and n-vertex endpoint labeled $n + 1$ in the image labeling, so condition (2) of Definition 4.10 holds for the new n-label $n + 1$ and all p-labels i as well. Hence the image labeling is a valid $(p, n + 1)$-simple-labeling, establishing (2).

For (3), note that two p-vertices have the same label in the multi-labeling if and only if they have the same label in the image simple labeling, so equivalence of p-labels is preserved under the map. Furthermore, equivalent multi-labelings have the same good pairs of n-vertices and the same good single n-vertices, and the good single n-vertices are divided into the same sets of vertices that share a common label. Hence equivalence of n-labels is also preserved under the map, so (3) holds. □

Definition A.14. *Let $\mathcal{C}$ and $\tilde{\mathcal{C}}$ be the set of all multi-labeling equivalence classes and simple-labeling equivalence classes, respectively, of an l-graph. For $\mathcal{L} \in \mathcal{C}$ and any multi-labeling in class $\mathcal{L}$, let $\tilde{\mathcal{L}} \in \tilde{\mathcal{C}}$ be the equivalence class of its image simple-labeling under the label simplifying map of Definition A.11. Then define $\varphi : \mathcal{C} \to \tilde{\mathcal{C}}$ by $\varphi(\mathcal{L}) = \tilde{\mathcal{L}}$.*

By parts (2) and (3) of Lemma A.13, the above construction of φ is well-defined. The remainder of this appendix shows that φ satisfies the conditions of Proposition 4.15.

Lemma A.15. *Suppose a multi-labeling of an l-graph has excess Δ. Then at most 42Δ pairs of consecutive p-vertices are such that their pair of p-labels appears (in some order) on three or more pairs of consecutive p-vertices.*

Proof. We induct on l. For $l = 2$ and 3, there are zero pairs of consecutive p-vertices satisfying the condition of the lemma, and $\Delta \geq 0$ by Lemma 4.4, so the result holds.

Suppose by induction that the result holds for $l-2$ and $l-1$, and consider a multi-labeling of an l-graph with $l \geq 4$. Let $d_1, \ldots, d_l$ be the number of n-labels on the first through l^{th} n-vertices, respectively, and let m be the number of distinct p-labels and n-labels. First suppose each distinct p-label which appears in the labeling appears at least twice. Then there are at most $\frac{l}{2}$ distinct p-labels. Lemma A.4 implies that there are at least $l - 6\Delta$ n-vertices such that all of its n-labels appear exactly twice. For any such n-vertex W and n-label j for W, consider the other n-vertex W' also having n-label j. The two p-vertices preceding and following W must have the same pair of labels as the two p-vertices preceding and following W', by conditions (1) and (3) of Definition 4.3. This implies that there are at most 6Δ pairs of p-labels which appear (in some order) on only one consecutive pair of p-vertices.

On the other hand, the number of distinct p-labels in the multi-labeling is at most one more than the number of distinct pairs of p-labels appearing on pairs of consecutive p-vertices. This is easily seen by considering the undirected graph with vertices $\{1, \ldots, p\}$, having an edge between $i, i' \in \{1, \ldots, p\}$ if and only if some consecutive pair of p-vertices have labels i and i'. The edges of this graph must form a single connected component, so the number of vertices adjacent to at least one edge (which is the number of distinct p-labels appearing in the multi-labeling) is at most one more than the number of edges. Note that by Lemma A.1, there are at most $\frac{\sum_{s=1}^{l} d_s}{2}$ distinct n-labels in the multi-labeling, so there are at least $m - \frac{\sum_{s=1}^{l} d_s}{2}$ distinct p-labels. Hence, there are at least $m - \frac{\sum_{s=1}^{l} d_s}{2} - 1 = \frac{l}{2} - \Delta$ distinct pairs of consecutive p-labels. By our previous argument, at least $\frac{l}{2} - 7\Delta$ of these pairs appear on at least two pairs of consecutive p-vertices. If c distinct pairs of p-labels appear on exactly two pairs of consecutive p-vertices, then by a pigeonhole argument, $2c + 3\left(\frac{l}{2} - 7\Delta - c\right) \leq l$, so $c \geq \frac{l}{2} - 21\Delta$. These account for at least $l - 42\Delta$ pairs of consecutive p-vertices, implying that at most 42Δ pairs of consecutive p-vertices have a pair of p-labels appearing three or more times. This establishes the result in this case.

Now suppose that there is some p-vertex, V, whose p-label appears only once in the labeling. Consider the $(l-1)$-graph or $(l-2)$-graph obtained by following either procedure (1) or procedure (2) of Lemma A.3. Note that this $(l-1)$-graph or $(l-2)$-graph has the same number of pairs of consecutive p-vertices such that their pair of p-labels appears on three or more pairs of consecutive p-vertices as in the original l-graph, since no such pair in the original l-graph could contain the p-label of V. On the other hand, our proof of Lemma 4.4 shows that this $(l-1)$-graph or $(l-2)$-graph has excess less than or equal to the excess of the original l-graph. Then the desired result follows from the induction hypothesis. □

The next lemma represents a key insight into the structure of the multi-labelings defined in Definition 4.3. It indicates that in any multi-labeling with small excess Δ, most of the non-single n-vertices must belong to a good pair, and in particular, if $\Delta = 0$, then all non-single n-vertices belong to good pairs.

Lemma A.16. *Suppose a multi-labeling of an l-graph has excess Δ and k single n-vertices. Then there are at least $\frac{l-k}{2} - 48\Delta$ good pairs of n-vertices.*

Proof. Let the multi-labeling have $d_1, \ldots, d_l$ n-labels for the first through l^{th} n-vertices, respectively, and m total distinct p-labels and n-labels. We induct on l. If $l = 2$, then Lemma A.2 implies $d_1 = d_2$, $m = d_1 + 2$, and $\Delta = 0$. If $d_1 = d_2 = 1$, then $k = 2$ and there are no good pairs, and if $d_1 = d_2 \geq 2$, then $k = 0$ and there is one good pair. Hence the result holds. If $l = 3$, then Lemma A.2 implies $d_1 = d_2 = d_3$, $m = d_1 + 3$, and $\Delta = \frac{d_1 - 1}{2}$. If $d_1 = d_2 = d_3 = 1$, then $k = 3$, $\Delta = 0$, and there are no good pairs. If $d_1 = d_2 = d_3 \geq 2$, then $k = 0$, $\Delta \geq \frac{1}{2}$, and there are still no good pairs. In either case, the result also holds.

Assume by induction that the result holds for $l-2$ and $l-1$, and consider a multi-labeling of an l-graph with $l \geq 4$. First suppose each distinct p-label which appears in the labeling appears at least twice. Then there are at most $\frac{l}{2}$ distinct p-labels. As there are $l-k$ non-single n-vertices, by Lemmas A.1 and A.4, there are at least $l-k-6\Delta$ non-single n-vertices such that each n-label of that vertex appears exactly twice. Let V be one such n-vertex. Suppose that V has two n-labels j_1 and j_2 that occur on two different n-vertices W_1 and W_2 respectively, in addition to V. Then by conditions (1) and (3) of Definition 4.3, the three pairs of consecutive p-vertices around V, W_1, and W_2 must have the same pair of (distinct) p-labels. By Lemma A.15, there are at most 42Δ such n-vertices V. Now suppose that all n-labels of V reappear on a single other n-vertex W_1, but W_1 has some additional n-label j not appearing on V. Then either all additional n-labels of W_1 appear at least three times, or there is some n-label j appearing on W_1 and a single other n-vertex W_2, but not on V. In the former case, the number of such vertices W_1 is at most 6Δ by Lemma A.4. As V is the unique n-vertex sharing an n-label with W_1 that appears exactly twice, this implies the number of such vertices V is also at most 6Δ. In the latter case, the three pairs of p-vertices around V, W_1, and W_2 must have the same pair of p-labels, so by Lemma A.15, the number of such vertices V is at most 42Δ. Combining the results from all of these cases, there are then at least $l-k-96\Delta$ non-single n-vertices V whose labels all appear exactly twice, on one other n-vertex V', and such that V' has no additional n-labels. These pairs (V, V') form at least $\frac{l-k}{2} - 48\Delta$ good pairs, so the conclusion holds.

Now suppose there is some p-vertex, V, whose p-label appears only once in the labeling. Let the n-vertex preceding V be U, the p-vertex preceding U be T, the n-vertex following V be W, and the p-vertex following W be X. By conditions (2) and (3) of Definition 4.3 and the fact that the p-vertex of V appears only once, U and W must have the same tuple of n-labels, up to reordering. Consider four cases:

(1) T and X have different p-labels, and U and W are single.
(2) T and X have different p-labels, and U and W each have $d \geq 2$ n-labels.
(3) T and X have the same p-label, and U and W are single.
(4) T and X have the same p-label, and U and W each have $d \geq 2$ n-labels.

In cases (1) and (2), remove V and W, and connect U to X. Lemma A.3 implies that the resulting graph is an $(l-1)$-graph with a valid multi-labeling. In case (1), this $(l-1)$-graph has $k-1$ single n-vertices, $\sum_{s=1}^{l} d_s - 1$ total n-labels, and $m-1$ total distinct p-labels and n-labels. Then by the induction hypothesis, it has at least $\frac{(l-1)-(k-1)}{2} - 48\left(\frac{(l-1)+(\sum_{s=1}^{l} d_s - 1)}{2} + 1 - (m-1)\right) = \frac{l-k}{2} - 48\Delta$ good pairs. Note that in case (1), this $(l-1)$-graph must have the same number of good pairs as the original l-graph, so the desired result holds.

In case (2), the $(l-1)$-graph has k single n-vertices, $\sum_{s=1}^{l} d_s - d$ total n-labels, and $m-1$ distinct p-labels and n-labels. By the induction hypothesis, as $d \geq 2$ by assumption, it has at least $\frac{(l-1)-k}{2} - 48\left(\frac{(l-1)+(\sum_{s=1}^{l} d_s - d)}{2} + 1 - (m-1)\right) > \frac{l-k}{2} - 48\Delta + 1$ good pairs. This $(l-1)$-graph can have at most one more good pair than the original l-graph. (It has exactly one more good pair if the removed n-vertex W had a tuple of n-labels that occurred exactly three times on three different n-vertices in the original l-graph.) So the number of good pairs in the original l-graph is at least $\frac{l-k}{2} - 48\Delta$ and the conclusion holds in this case as well.

In cases (3) and (4), remove U, V, W, and X, and connect T to the n-vertex after X. Lemma A.3 implies that the resulting graph is an $(l-2)$-graph with a valid multi-labeling. In case (3), this $(l-2)$-graph has $k-2$ single n-vertices, $\sum_{s=1}^{l} d_s - 2$ total n-labels, and either $m-2$ distinct p-labels and n-labels if the removed n-vertices had an n-label appearing only those two times, or $m-1$ distinct p-labels and n-labels otherwise. Suppose the former. Then, by the induction hypothesis, this $(l-2)$-graph has at least $\frac{(l-2)-(k-2)}{2} - 48\left(\frac{(l-2)+(\sum_{s=1}^{l} d_s - 2)}{2} + 1 - (m-2)\right) = \frac{l-k}{2} - 48\Delta$ good

pairs, and it has the same number of good pairs as the original l-graph. If, instead, the $(l-2)$-graph has $m-1$ distinct p-labels and n-labels, then it can have at most one more good pair than the original l-graph. (It has exactly one more good pair if the $(l-2)$-graph has a pair of n-vertices having the n-label of the removed vertices U and W, and this pair now forms a good pair.) But in this case, the $(l-2)$-graph has at least $\frac{(l-2)-(k-2)}{2} - 48\left(\frac{(l-2)+(\sum_{s=1}^{l} d_s - 2)}{2} + 1 - (m-1)\right) > \frac{l-k}{2} - 48\Delta + 1$ good pairs, so the original l-graph must have at least $\frac{l-k}{2} - 48\Delta$ good pairs in this case as well.

Finally, in case (4), the $(l-2)$-graph has k single n-vertices, $\sum_{s=1}^{l} d_s - 2d$ total n-labels, and at least $m-d-1$ and at most $m-1$ distinct p-labels and n-labels. If it has exactly $m-d-1$ distinct p-labels and n-labels, then we must have removed a good pair, and by the induction hypothesis, the $(l-2)$-graph has at least $\frac{(l-2)-k}{2} - 48\left(\frac{(l-2)+(\sum_{s=1}^{l} d_s - 2d)}{2} + 1 - (m-d-1)\right) = \frac{l-k}{2} - 48\Delta - 1$ good pairs. Hence the original l-graph had at least $\frac{l-k}{2} - 48\Delta$ good pairs. If, instead, the $(l-2)$-graph has $m-c-1$ distinct p-labels and n-labels for $0 \leq c < d$ (so that $d-c$ distinct n-labels in the removed pair of n-vertices U and W appear more than just those two times), then U and W cannot be a good pair in the original l-graph, and the $(l-2)$-graph can have at most $d-c$ more good pairs than the l-graph, one for each distinct n-vertex label of U and W that appeared more than twice in the l-graph. The $(l-2)$-graph has at least $\frac{(l-2)-k}{2} - 48\left(\frac{(l-2)+(\sum_{s=1}^{l} d_s - 2d)}{2} + 1 - (m-c-1)\right) > \frac{l-k}{2} - 48\Delta + d - c$ good pairs, which implies that the original l-graph had at least $\frac{l-k}{2} - 48\Delta$ good pairs.

This completes the induction in all cases, so the conclusion holds for all l. □

Remark A.17. *In the context of Theorem 2.5 and Lemma 4.7, if $a = 0$, then only multi-labeling equivalence classes with no single n-vertices contribute to the sum in eq. (6). By Lemma A.16, the multi-labelings with no single n-vertices and excess $\Delta = 0$, which comprise the dominant term of this sum, must be such that all n-vertices belong to good pairs. Then there are exactly $\frac{\sum_{s=1}^{l} d_s}{2}$ distinct n-labels in such multi-labelings, which in turn implies by the definition of Δ that there are exactly $\frac{l}{2} + 1$ distinct p-labels. By Remark A.9, this identifies the sequence of p-labels appearing along the l-graph cycle as a traversal of a tree with $\frac{l}{2}$ edges in the complete graph having vertices $\{1, \ldots, p\}$, where each edge of the tree is traversed exactly once in each direction. This corresponds exactly to Wigner paths in the combinatorial proof of the semicircle law for Wigner matrices and explains the emergence of the semicircle law as the limit measure $\mu_{a,\nu,\gamma}$ in Theorem 2.5 in the case where $a = 0$.*

Remark A.18. *A statement analogous to Lemma A.16 for the single n-vertices does not hold, i.e., it is not true in general that if a multi-labeling of an l-graph has excess Δ and k single n-vertices, then at least $k - O(\Delta)$ of these are good singles.*

Lemma A.19. *Suppose a multi-labeling of an l-graph has excess Δ. Then there are at most 2Δ good pairs of n-vertices such that the two vertices in the pair are consecutive in the l-graph cycle and the p-label of the p-vertex between them appears at least twice in the labeling.*

Proof. Suppose that (V, V') is such a pair, W is the p-vertex between them, and W has p-label i. If i appears on any p-vertex that is not between two consecutive n-vertices forming a good pair, then change the p-label of W to a new p-label not yet appearing in the multi-labeling, and do this for every such p-vertex W with label i (picking a different new p-label each time). If i only appears on p-vertices between consecutive n-vertices forming good pairs, and there are c such p-vertices including W, then change the p-labels of $c-1$ of these p-vertices to $c-1$ new labels not yet appearing in the multi-labeling. Note that changing the p-label of any p-vertex between two consecutive n-vertices forming a good pair to a new p-label not yet appearing in the labeling cannot violate any of the conditions of Definition 4.3, so the resulting labeling is still a valid multi-labeling. If x is the number of good pairs satisfying the condition of the lemma, then we have added at least $\frac{x}{2}$ distinct new p-labels to the multi-labeling. If there were originally m distinct n-labels

and p-labels and $d_1, \ldots, d_l$ n-labels on the first through l^{th} n-vertices, respectively, then Lemma 4.4 implies $m + \frac{x}{2} \leq \frac{l+\sum_{s=l} d_s}{2} + 1$, so $x \leq 2\Delta$. □

Definition A.20. *In a multi-labeling of an l-graph, a distinct p-label i that appears in the multi-labeling is a* **connector** *if, among all n-vertices that are adjacent to any p-vertex with label i, exactly two of them are bad singles and the remainder of them are either good singles or part of good pairs. Two bad single n-vertices are* **connected** *if they are these two n-vertices corresponding to a connector i. A sequence of bad single n-vertices $W_1, \ldots, W_a$ is a* **connected cycle** *if W_1 is connected to W_2, W_2 is connected to W_3, etc., and W_a is connected to W_1.*

Note that in the above definition, "connector" refers to a distinct p-label i, not to any specific p-vertex having label i, and any two "connected" bad single n-vertices are adjacent to p-vertices having some connector label i but these p-vertices may be distinct vertices in the l-graph. Each bad single n-vertex may be connected to at most two other bad single n-vertices (where the connectors are the p-labels of the p-vertices adjacent to that bad single n-vertex), and hence this notion of connectedness partitions the set of bad single n-vertices into connected components that are either individual vertices, linear chains, or cycles. The motivation for the above definition comes from the observation that if two bad single n-vertices are connected, then they must have the same n-label, as follows from condition (3) of Definition 4.3 and the fact that n-labels appearing on good singles and good pairs must be distinct from those appearing on n-vertices that are not good singles nor good pairs.

Lemma A.21. *Suppose a multi-labeling of an l-graph has excess Δ and k single n-vertices, of which k' are good single and $k - k'$ are bad single. Then at least $k - k' - (288D + 2)\Delta$ distinct p-labels are connectors, and there are at most $(192D+1)\Delta$ connected cycles of bad single n-vertices.*

Proof. Suppose the multi-labeling is a (p, n)-multi-labeling. Construct an undirected multi-graph G on p vertices labeled $\{1, \ldots, p\}$, with each edge of G having one label in $\{1, \ldots, n\}$, such that the following is true: Corresponding to each n-label j of each n-vertex V in the multi-labeling, if V is preceded and followed by p-vertices having labels i_1 and i_2, there is an edge between vertices i_1 and i_2 of G with label j. (Hence, if the multi-labeling has $\sum_{s=1}^{l} d_s$ total n-labels, then G has $\sum_{s=1}^{l} d_s$ total edges.) Condition (3) of Definition 4.3 states that for any n-label j, each vertex of G has even degree in the sub-graph consisting of only edges with label j.

We will sequentially remove the edges of G corresponding to the good pair and good single n-vertices of the multi-labeled l-graph, until only the edges of G corresponding to the n-vertices that are not good pairs or good singles remain. At any stage of this removal process, let us call a vertex of G "active" if there is at least one edge still adjacent to that vertex. Let us define a "component" as the set of active vertices that may be reached by traversing the remaining edges of G from a particular active vertex. (Hence a component of G is a connected component, in the standard sense, that contains at least two vertices.) We will keep track of the quantity

$$M = \#\{\text{active vertices}\} + \#\{\text{distinct edge labels}\} - \#\{\text{components}\}.$$

Note that initially, if the l-graph multi-labeling has m distinct p-labels and n-labels, then m is also the number of active vertices plus the number of distinct edge labels of G. Also, G initially has only one component, so $M = m - 1$. Let us now remove the edges of G corresponding to the good pairs of the l-graph. If an n-vertex of a good pair has d n-labels, then the good pair corresponds to $2d$ edges between a single pair of vertices in G, having d distinct edge labels that do not occur elsewhere in G. Hence removing these $2d$ edges of G removes d distinct edge labels, and if this also changes the connectivity structure of G, then either the number of components increases by 1, the number of components stays the same but the number of active vertices decreases by 1, or the number of components decreases by 1 and the number of active vertices decreases by 2. In all of these cases, upon removing these $2d$ edges from G, M decreases by at most $d + 1$. Then after

removing all edges of G corresponding to good pairs, $M \geq m-1-\left(\frac{\sum_{s=1}^{l} d_s - k}{2}\right) - \left(\frac{l-k}{2}\right) = k - \Delta$, as there are at most $\frac{\sum_{s=1}^{l} d_s - k}{2}$ distinct n-vertex labels for the good pairs and at most $\frac{l-k}{2}$ good pairs.

Let us now remove the edges of G corresponding to the good single n-vertices in the l-graph. Let j be an n-label that appears on a good single n-vertex, and consider removing the edges of G with label j one at a time. As each vertex of G has even degree in the subgraph of edges of G with label j, when the first such edge is removed, the number of components and active vertices cannot change. Subsequently, the removal of each additional edge might increase the number of components by 1, keep the number of components the same and decrease the number of active vertices by 1, or decrease the number of components by 1 and the number of active vertices by 2. When the last such edge is removed, there are no longer any edges with label j by the definition of a good single, so the number of distinct edge labels decreases by 1. Hence removing all edges with label j decreases M by at most the number of such edges, and $M \geq k - k' - \Delta$ after removing the edges corresponding to all k' good singles.

Call the resulting graph G'. Note that every vertex of G' still has even degree in the sub-graph consisting of edges with label j, for any j, and in particular, every active vertex of G' has degree at least two. Then by Definition A.20, a p-label i of the l-graph is a connector if and only if vertex i has degree exactly two in G', in which case the edges adjacent to i in G' must have the same label j, and the n-vertices with label j in the l-graph are the bad singles connected by connector i. A connected cycle of bad single n-vertices in the l-graph corresponds to the edges of a cycle of (necessarily distinct) vertices in G' with degree exactly two.

The number of distinct edge labels that remain in G' is the number of distinct n-labels in the original l-graph multi-labeling that appear on n-vertices that are not good singles nor part of good pairs, which by Lemma A.16 is at most $96D\Delta$. Hence the number of active vertices minus the number of components of G' is at least $k-k'-(96D+1)\Delta$, by our lower bound on M. The number of total edges in G' is at most $k - k' + 96D\Delta$, with $k - k'$ of them corresponding to bad single n-vertices of the l-graph and at most $96D\Delta$ of them corresponding to non-single n-vertices that are not part of good pairs. Then the total vertex degree of G' is at most $2(k - k' + 96D\Delta)$. As each active vertex in G' has degree at least two, this implies there are at most $k-k'+96D\Delta$ active vertices. Then there are at most $(192D+1)\Delta$ components in G', and hence at most $(192D+1)\Delta$ connected cycles of bad single n-vertices in the l-graph. Furthermore, if x active vertices in G' have degree exactly two in G', then as there are at least $k-k'-(96D+1)\Delta$ active vertices, a pigeonhole argument implies $2x+4(k-k'-(96D+1)\Delta-x) \leq 2(k-k'+96D\Delta)$, so $x \geq k-k'-(288D+2)\Delta$. Hence there are at least $k - k' - (288D+2)\Delta$ connectors in the l-graph multi-labeling. □

Proof of Proposition 4.15. For notational convenience, let C denote a positive constant that may depend on D and that may change from instance to instance. Let φ be as defined in Definition A.14. As φ preserves the p-labels, up to reordering, clearly condition (1) of Proposition 4.15 holds.

To verify condition (2), let $\mathcal{L} \in \mathcal{C}$ be any multi-labeling equivalence class. Let $\varphi(\mathcal{L})$ have $\tilde{k}$ n-vertices with non-empty label. This means $\mathcal{L}$ has $\tilde{k}$ n-vertices that do not belong to a good pair. These vertices have at least $\tilde{k}$ total n-labels in $\mathcal{L}$, implying that there are at most $\sum_{s=1}^{l} d_s - \tilde{k}$ total n-labels on the good pair vertices, where $d_1, \ldots, d_l$ are the number of n-labels on the first through l^{th} n-vertices, respectively, in $\mathcal{L}$. These good pair vertices account for at most $\frac{\sum_{s=1}^{l} d_s - \tilde{k}}{2}$ distinct n-labels in $\mathcal{L}$, and these are mapped to the empty label under the label-simplifying map. Furthermore, by Lemma A.16, there are at most $2C\Delta(\mathcal{L})$ n-vertices that are not single and that also do not belong to a good pair in $\mathcal{L}$, and hence these have at most $2CD\Delta(\mathcal{L})$ additional distinct n-labels in $\mathcal{L}$ that are mapped to the new n-label $n+1$ under the label-simplifying map. Note that any bad single n-vertex has an n-label that is the same as one of these $2CD\Delta(\mathcal{L})$ distinct n-labels (otherwise it is a good single by definition), and the n-label of any good single n-vertex is preserved under the label-simplifying map. Hence, if m is the number of total distinct p-labels

and n-labels in $\mathcal{L}$ and $\tilde{m}$ is the number of total distinct p-labels and non-empty n-labels in $\varphi(\mathcal{L})$, then, as the set of distinct p-labels is unchanged under the label-simplifying map, this implies $\tilde{m} \geq m - \frac{\sum_{s=1}^{l} d_s - \tilde{k}}{2} - 2CD\Delta(\mathcal{L})$, so $\tilde{\Delta}(\varphi(\mathcal{L})) = \frac{l+\tilde{k}}{2} + 1 - \tilde{m} \leq (2CD+1)\Delta(\mathcal{L})$. Hence condition (2) holds.

It remains to verify condition (3). Fix $\tilde{\mathcal{L}} \in \tilde{\mathcal{C}}$, and let the "canonical simple-labeling" in the class $\tilde{\mathcal{L}}$ be the one in which each ith new p-label that appears in the l-graph cycle is label i and each j^{th} new n-label that appears in the l-graph cycle is label j. Note that as there are at most l distinct p-labels and l distinct n-labels, the canonical simple-labeling is an (l,l)-simple-labeling of the l-graph. For notational convenience, let us still denote this canonical simple-labeling by $\tilde{\mathcal{L}}$ when the meaning is clear. Consider the below (non-determined) procedure that constructs a multi-labeling from $\tilde{\mathcal{L}}$.

(1) Choose an n-label in $\{1,\ldots,l\}$ to be the "new label", or assume there is no new label. (n-vertices in $\tilde{\mathcal{L}}$ with the new label will be the ones that are neither good singles nor part of good pairs in the multi-labeling.)
(2) For all p-vertices, copy its p-label from $\tilde{\mathcal{L}}$ to the multi-labeling, and for all n-vertices with non-empty n-label that is not the "new label" in $\tilde{\mathcal{L}}$, copy its n-label from $\tilde{\mathcal{L}}$ to the multi-labeling.
(3) Among n-vertices having the new label in $\tilde{\mathcal{L}}$, choose a subset S of them that will correspond to the non-single n-vertices that are not part of good pairs.
(4) For each n-vertex in S, choose the size of its n-label tuple in the multi-labeling to be between 2 and D, and pick the n-labels for that tuple.
(5) For each n-vertex with the new label but not belonging to S, pick one n-label for that n-vertex in the multi-labeling.
(6) For all n-vertices with empty label in $\tilde{\mathcal{L}}$, pair them up into good pairs for the multi-labeling.
(7) Let $\mathcal{G}$ be the set of good pairs (V,V') in the multi-labeling that are consecutive n-vertices in the l-graph, and such that the p-label of the p-vertex between them appears at least twice. Choose an ordered subset of $\mathcal{G}$. For each (V,V') in this ordered subset, if W is the p-vertex between V and V', choose some other p-vertex W' having the same p-label as W, and either reverse the sequence of vertices from W to W' or reverse the sequence of vertices from W' to W.
(8) For each good pair in the multi-labeling, choose the size of its n-label tuple to be between 2 and D, pick the n-labels for the first vertex of the good pair, and pick a permutation of these n-labels for the second vertex of the good pair.

The above procedure is non-determined in the sense that there are many ways to perform each of the above steps, and hence many different multi-labelings may be the output of the procedure for a single canonical simple-labeling $\tilde{\mathcal{L}}$. (The above procedure may, in addition, construct labelings that are not valid multi-labelings according to Definition 4.3, but those will be irrelevant for our argument.) We claim that for any $\mathcal{L} \in \varphi^{-1}(\tilde{\mathcal{L}})$, there is a multi-labeling in class $\mathcal{L}$ that may be constructed from $\tilde{\mathcal{L}}$ according to the above procedure, and furthermore that this multi-labeling is a (l, Dl)-multi-labeling such that, for each good pair, the n-labels of the first n-vertex in the pair are the smallest n-labels not yet appearing in the labeling and are in sorted order.

To verify this claim, note that the steps of the above procedure "invert" the label-simplifying map in Definition A.11. The "new label" chosen in step (1) above corresponds to the label $n+1$ that is given to n-vertices that are neither good singles nor part of good pairs in the multi-labeling under the label-simplifying map (except, of course, if we take $\tilde{\mathcal{L}}$ to be the canonical simple-labeling equivalent to the output of the label-simplifying map, then the new label is no longer $n+1$ in $\tilde{\mathcal{L}}$). If no new label is chosen, this implies that the multi-labeling we construct has all of its n-vertices being either a good single or part of a good pair. Steps (2)–(6) and (8) above invert the process by which p-labels and n-labels are mapped from the multi-labeling to the simple-labeling under steps (2) and

(3) of Definition A.11 for the label-simplifying map. Note that if step (1) of the label-simplifying map in Definition A.11 is performed on any multi-labeling, each reversal that is performed causes an additional good pair (V, V') of n-vertices that were not consecutive to become consecutive in the l-graph, and they remain consecutive under each subsequent reversal. Hence, starting with the final multi-labeling in which all good pairs are proper, we may perform this sequence of reversals in reverse order to recover the original multi-labeling. So step (7) above inverts step (1) of Definition A.11. This establishes that, for any multi-labeling equivalence class $\mathcal{L} \in \varphi^{-1}(\tilde{\mathcal{L}})$, there exists some multi-labeling L in class $\mathcal{L}$ that may be constructed by the above procedure from $\tilde{\mathcal{L}}$. The above procedure picks new n-labels for the non-good-single n-vertices of L in steps (4), (5), and (8), but since it does not specify which n-labels are picked, there is a (l, Dl)-multi-labeling equivalent to L that may also be constructed by the above procedure, and such that for each of its good pairs, the n-labels of the first n-vertex in the pair are the smallest n-labels not yet appearing in the labeling and are in sorted order. This establishes our claim.

Thus, to verify condition (3) of the proposition, it suffices to upper-bound the number of ways in which each of the above steps (1)–(8) may be performed while still ensuring that the resulting multi-labeling is a valid (l, Dl)-multi-labeling satisfying Definition 4.3, having excess Δ_0, and such that the first n-vertex of each good pair has n-labels that are the smallest n-labels not yet appearing in the labeling and are in sorted order. As there are at most l distinct n-labels in $\tilde{\mathcal{L}}$, there are at most $l + 1$ ways of choosing the new label or choosing no new label in step (1). There is only one way of performing step (2). By Lemma A.16, for a multi-labeling with excess Δ_0, there can be at most $C\Delta_0$ n-vertices that are not single and also do not belong to a good pair. Hence, to obtain a multi-labeling with excess Δ_0, we may restrict our selection of S in step (3) to be of size at most $C\Delta_0$, so there are at most $(l+1)^{C\Delta_0}$ ways of choosing S in step (3). To perform step (4), for each vertex in S, we may first choose the number of n-labels d between 1 and D, and then there are at most $(Dl)^d$ ways of choosing the n-labels for that vertex.

For step (5), suppose that k single n-vertices and k' good single n-vertices have been identified in steps (1)–(4). In step (5) we must assign n-labels to each of the $k - k'$ n-vertices that are bad singles. Recall the notions of connectors and connected bad singles from Definition A.20. By Lemma A.21, there are at least $k - k' - C\Delta_0$ connectors. By condition (3) of Definition 4.3, any two of the $k - k'$ bad single n-vertices connected by a connector must be given the same n-label in the multi-labeling. Hence, going through the connectors one-by-one, each successive connector constrains the n-label of one more bad single n-vertex, unless that connector closes a connected cycle of such vertices. But as there are at most $C\Delta_0$ total connected cycles by Lemma A.21, this implies that the number of bad single n-vertices that we can freely label at most $C\Delta_0$ (rather than the naive bound of at most $k - k'$). Then there are at most $(Dl)^{C\Delta_0}$ ways to perform step (5).

For step (6), recall from Remark A.9 that the pairs of p-vertices surrounding the two n-vertices of each good pair must have the same pair of p-labels. By Lemma A.15, for all but at most $C\Delta_0$ of the n-vertices with empty label, this pairing is uniquely determined. Then there are at most $(C\Delta_0)^{C\Delta_0}$ ways of performing the pairing in step (6). For step (7), Lemma A.19 shows that $|\mathcal{G}| \leq 2\Delta_0$. Then there are at most $2\Delta_0$ ways of choosing each successive element in the ordered subset of $\mathcal{G}$, and at most $2l$ ways of choosing the other p-vertex W' as well as which half of the cycle to reverse for each such added element. As the size of the ordered subset is also at most $2\Delta_0$, and we may choose to not add any more elements to the ordered subset at any point, there are at most $(4\Delta_0 l+1)^{2\Delta_0}$ ways of performing step (7). Finally, for step (8), for each good pair we may first choose the number of n-labels d between 2 and D. As we are requiring that the first vertex of the pair have the smallest n-labels not yet appearing in the multi-labeling and in sorted order, this determines uniquely the choice of n-labels for this first vertex in the good pair. We may then choose the permutation of these labels for the second vertex in the pair from one of $d!$ choices.

Combining the above arguments, we obtain the bound

$$\sum_{\substack{\mathcal{L}\in\varphi^{-1}(\tilde{\mathcal{L}})\\ \Delta(\mathcal{L})=\Delta_0}}\prod_{s=1}^{l}\frac{|a_{d_s}(\mathcal{L})|}{(d_s(\mathcal{L})!)^{1/2}} \le (l+1)\sum_S\left(|a_1|^{\tilde{k}(\tilde{\mathcal{L}})-|S|}\left(\sum_{d=2}^{D}(Dl)^d\frac{|a_d|}{(d!)^{1/2}}\right)^{|S|}(Dl)^{C\Delta_0}(C\Delta_0)^{C\Delta_0}\right.$$
$$\left.(4\Delta_0 l+1)^{2\Delta_0}\left(\sum_{d=2}^{D}d!\frac{a_d^2}{d!}\right)^{\frac{l-\tilde{k}(\tilde{\mathcal{L}})}{2}}\right)$$
$$\le (l+1)(Cl)^{C\Delta_0}|a|^{\tilde{k}(\tilde{\mathcal{L}})}(\nu-a^2)^{\frac{l-\tilde{k}(\tilde{\mathcal{L}})}{2}}\sum_S|a|^{-|S|}\left(\sum_{d=2}^{D}(Dl)^d\frac{|a_d|}{(d!)^{1/2}}\right)^{|S|},$$

where the summation over S represents the sum over all possible sets S selected by step (3) of the procedure above. As $|S|\le C\Delta_0$ by our preceding argument, this implies

$$|a|^{-|S|}\left(\sum_{d=2}^{D}(Dl)^d\frac{|a_d|}{(d!)^{1/2}}\right)^{|S|}\le(Cl)^{C\Delta_0}|a|^{-|S|}\sqrt{D}\left(\sum_{d=2}^{D}\frac{a_d^2}{d!}\right)^{\frac{|S|}{2}}\le\sqrt{D}(Cl)^{C\Delta_0}\left(\frac{\sqrt{\nu}}{|a|}\right)^{C\Delta_0}.$$

The sum is over at most $(l+1)^{C\Delta_0}$ possible sets S, so this verifies condition (3) of the proposition upon noting that $\sqrt{D}(l+1)^{C\Delta_0+1}(Cl)^{C\Delta_0}\le l^{C_3+C_4\Delta_0}$ for some constants $C_3, C_4>0$ and all $l\ge 2$. □

Appendix B. Moments of a deformed GUE matrix

In this Appendix, we prove Proposition 4.9. Recall Definition 4.8 of $M_{\tilde{n},\tilde{p}}$, $W_{\tilde{p}}$, $V_{\tilde{n},\tilde{p}}$, and $Z_{\tilde{n},\tilde{p}}$. Throughout this section, we will use p and n in place of $\tilde{p}$ and $\tilde{n}$, and we will suppress the dependence of W, V, and Z on n and p.

Lemma B.1. *Suppose $n,p\to\infty$ with $\frac{p}{n}\to\gamma$. Then $\|M_{n,p}\|\to\|\mu_{a,\nu,\gamma}\|$ almost surely*

Proof. Recall $M_{n,p}=\sqrt{\frac{\gamma(\nu-a^2)}{p}}W+\frac{a}{n}V$, where $V=ZZ^T-D$ and $D=\mathrm{diag}(\|Z_i\|_2^2)$. We will apply Proposition 8.1 of [6], conditional on V. The empirical spectral distribution of $\frac{1}{n}ZZ^T$ converges weakly a.s. to $\mu_{MP,\gamma}$, the Marcenko-Pastur law with parameter γ. By a standard chi-squared tail bound and a union bound, for any $\varepsilon>0$, $\mathbb{P}[\max_{1\le i\le p}|\|Z_i\|_2^2-n|>\varepsilon n]\le 2pe^{-\frac{n\varepsilon^2}{8}}$. Then the Borel-Cantelli lemma implies $\left\|\frac{1}{n}D-I\right\|\to 0$ a.s., and hence the empirical spectral distribution of $\frac{a}{n}V$ converges weakly a.s. to the translated and scaled Marcenko-Pastur law $\mu_{MP,shift}$ of Proposition 2.11. Furthermore, $\mu_{MP,\gamma}$ has compact support, and the maximal distance between an eigenvalue of $\frac{1}{n}ZZ^T$ and the support of $\mu_{MP,\gamma}$ converges to 0 a.s. by the results of [34] and [1]. Hence the same is true of $\mu_{MP,shift}$ and $\frac{a}{n}V$.

Let $V=O\Lambda O^T$ where O is the real orthogonal matrix that diagonalizes V. Then the spectrum of $M_{n,p}$ is the same as that of $\sqrt{\frac{\gamma(\nu-a^2)}{p}}O^TWO+\frac{a}{n}\Lambda$. By the above argument, conditional on V, $\frac{a}{n}\Lambda$ is a non-random diagonal matrix whose empirical spectral distribution converges weakly to $\mu_{MP,shift}$ and such that the maximal distance between any of its diagonal entries and $\mathrm{supp}(\mu_{MP,shift})$ converges to 0 (a.s. in V). Furthermore, O^TWO is still distributed as the GUE by unitary invariance. Hence the conditions of [6] are satisfied with no spike eigenvalues, so, conditional on V, Proposition 8.1 of [6] implies that the spectral norm of $\sqrt{\frac{\gamma(\nu-a^2)}{p}}O^TWO+\frac{a}{n}\Lambda$ converges a.s. to $\sup\{|x|:x\in\mathrm{supp}(\mu_{sc}\boxplus\mu_{MP,shift})\}$. As this convergence holds a.s. in V and the limit does not depend on V, it holds a.s. unconditionally as well. The result then follows from Proposition 2.11. □

Lemma B.2. *Suppose $n, p \to \infty$ with $\frac{p}{n} \to \gamma$, let $l := l(n)$ be such that $\frac{l(n)}{n} \to 0$, and let $\mathcal{B}_n$ be any event. Then there exist positive constants $C := C_{a,\nu,\gamma}$ and $c := c_{a,\nu,\gamma}$ such that $\mathbb{E}[\|M_{n,p}\|^l \mathbb{1}\{\mathcal{B}_n\}] \leq C^l \mathbb{P}[\mathcal{B}_n] + e^{-cn}$ for all large n.*

Proof. Note

$$\|M_{n,p}\| \leq \sqrt{\frac{\gamma(\nu - a^2)}{p}}\|W\| + \frac{|a|}{n}\|ZZ^T\| + \frac{|a|}{n}\max_{1\leq i\leq p}\|Z_i\|_2^2.$$

By Corollary 2.3.5 of [29], there exist positive constants A and B such that for all $t \geq A$,

$$\mathbb{P}[\|W\| > t\sqrt{p}] \leq Ae^{-Btp}.$$

By Corollary 5.35 of [31], for all $t \geq 0$,

$$\mathbb{P}[\|ZZ^T\| > (\sqrt{n} + \sqrt{p} + \sqrt{tn})^2] \leq \mathbb{P}[\|Z\| > \sqrt{n} + \sqrt{p} + \sqrt{tn}] \leq 2e^{-\frac{tn}{2}}$$

By a standard chi-squared tail bound and a union bound, for all $t \geq 0$,

$$\mathbb{P}\left[\max_{1\leq i\leq p}\|Z_i\|_2^2 > n + \sqrt{t}n\right] \leq pe^{-\frac{tn}{8}}.$$

Hence, for all $t \geq A$ and sufficiently large n,

$$\mathbb{P}\left[\|M_{n,p}\| > t\sqrt{\gamma(\nu - a^2)} + |a|(1.1 + \sqrt{\gamma} + \sqrt{t})^2 + |a|(1 + \sqrt{t})\right] \leq Ae^{-Btp} + 2e^{-\frac{tn}{2}} + pe^{-\frac{tn}{8}}.$$

So there exist constants $C, \varepsilon > 0$ depending on a, ν, γ such that, for all $t \geq C$ and sufficiently large n, $\mathbb{P}[\|M_{n,p}\| > t] \leq e^{-\varepsilon tn}$. Then we may write

$$\begin{aligned}
\mathbb{E}\left[\|M_{n,p}\|^l \mathbb{1}\{\mathcal{B}_n\}\right] &= \mathbb{E}\left[\|M_{n,p}\|^l \mathbb{1}\{\mathcal{B}_n\}\mathbb{1}\{\|M_{n,p}\| \leq C\}\right] + \mathbb{E}\left[\|M_{n,p}\|^l \mathbb{1}\{\mathcal{B}_n\}\mathbb{1}\{\|M_{n,p}\| > C\}\right] \\
&\leq C^l\mathbb{P}[\mathcal{B}_n] + \int_{C^l}^{\infty} \mathbb{P}\left[\|M_{n,p}\|^l > t\right] dt \\
&= C^l\mathbb{P}[\mathcal{B}_n] + \int_{C}^{\infty} \mathbb{P}[\|M_{n,p}\| > s] \cdot ls^{l-1} ds \\
&\leq C^l\mathbb{P}[\mathcal{B}_n] + l\int_{C}^{\infty} e^{-\varepsilon sn + (l-1)\log s} ds \\
&\leq C^l\mathbb{P}[\mathcal{B}_n] + l\int_{C}^{\infty} e^{-(\varepsilon n - l)s} ds \\
&= C^l\mathbb{P}[\mathcal{B}_n] + \frac{l}{\varepsilon n - l} e^{-(\varepsilon n - l)C}
\end{aligned}$$

for all large n. As $l = o(n)$, the result follows upon setting $c = \frac{C\varepsilon}{2}$. □

Lemma B.3. *Suppose $n, p \to \infty$ with $\frac{p}{n} \to \gamma$. Then $\mathbb{E}[\|M_{n,p}\|] \to \|\mu_{a,\nu,\gamma}\|$.*

Proof. By Lemma B.1, $\|M_{n,p}\| \to \|\mu_{a,\nu,\gamma}\|$ almost surely. Then $\liminf \mathbb{E}[\|M_{n,p}\|] \geq \mathbb{E}[\liminf \|M_{n,p}\|] = \|\mu_{a,\nu,\gamma}\|$ by Fatou's lemma, so it suffices to show $\limsup \mathbb{E}[\|M_{n,p}\|] \leq \|\mu_{a,\nu,\gamma}\| + \varepsilon$ for all $\varepsilon > 0$. Let $\mathcal{B}_n = \left\{\|M_{n,p}\| > \|\mu_{a,\nu,\gamma}\| + \frac{\varepsilon}{2}\right\}$. Then

$$\mathbb{E}[\|M_{n,p}\|] = \mathbb{E}[\|M_{n,p}\|\mathbb{1}\{\mathcal{B}_n^C\}] + \mathbb{E}[\|M_{n,p}\|\mathbb{1}\{\mathcal{B}_n\}] \leq \|\mu_{a,\nu,\gamma}\| + \tfrac{\varepsilon}{2} + \mathbb{E}[\|M_{n,p}\|\mathbb{1}\{\mathcal{B}_n\}].$$

Lemma B.1 implies $\mathbb{P}[\mathcal{B}_n] \to 0$, so Lemma B.2 (with $l = 1$) implies $\mathbb{E}[\|M_{n,p}\|\mathbb{1}\{\mathcal{B}_n\}] \to 0$ as well. Then $\mathbb{E}[\|M_{n,p}\|] \leq \|\mu_{a,\nu,\gamma}\| + \varepsilon$ for all large n, as desired. □

Lemma B.4. *Suppose $F : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz on a set $G \subseteq \mathbb{R}^k$, i.e. $|F(x) - F(y)| \leq L\|x - y\|_2$ for all $x, y \in G$. Let $\xi \sim N(0, I_d)$. Then there exists a function $\tilde{F} : \mathbb{R}^d \to \mathbb{R}$ such that $\tilde{F}(x) = F(x)$ for all $x \in G$, $|\tilde{F}(x) - \tilde{F}(y)| \leq L\|x - y\|_2$ for all $x, y \in \mathbb{R}^k$, and, for all $\Delta > 0$,*

$$\mathbb{P}[F(\xi) - \mathbb{E}F(\xi) \geq \Delta + |\mathbb{E}F(\xi) - \mathbb{E}\tilde{F}(\xi)| \text{ and } \xi \in G] \leq e^{-\frac{\Delta^2}{2L^2}}.$$

Proof. Let $\tilde{F}(x) = \inf_{x' \in G}(F(x') + L\|x - x'\|_2)$. Note that if $x \in G$, then $F(x) \le F(x') + L\|x - x'\|_2$ for all $x' \in G$, so $\tilde{F}(x) = F(x)$. Also, for any $x, y \in \mathbb{R}^k$ and $\varepsilon > 0$, there exists $x' \in G$ such that $\tilde{F}(x) \ge F(x') + L\|x - x'\|_2 - \varepsilon$. Then by definition, $\tilde{F}(y) \le F(x') + L\|y - x'\|_2$, so $\tilde{F}(y) - \tilde{F}(x) \le L\|y - x'\|_2 - L\|x - x'\|_2 + \varepsilon \le L\|x - y\|_2 + \varepsilon$. Similarly, $\tilde{F}(x) - \tilde{F}(y) \le L\|x - y\|_2 + \varepsilon$. This holds for all $\varepsilon > 0$, so $|\tilde{F}(x) - \tilde{F}(y)| \le L\|x - y\|_2$. Finally, applying Gaussian concentration of measure for the Lipschitz function $\tilde{F}$,

$$\begin{aligned}
&\mathbb{P}[F(\xi) - \mathbb{E}F(\xi) \ge \Delta + |\mathbb{E}F(\xi) - \mathbb{E}\tilde{F}(\xi)| \text{ and } \xi \in G] \\
&\quad = \mathbb{P}[\tilde{F}(\xi) \ge \Delta + |\mathbb{E}F(\xi) - \mathbb{E}\tilde{F}(\xi)| + \mathbb{E}F(\xi) \text{ and } \xi \in G] \\
&\quad \le \mathbb{P}[\tilde{F}(\xi) \ge \Delta + \mathbb{E}\tilde{F}(\xi)] \le e^{-\frac{\Delta^2}{2L^2}}.
\end{aligned}$$

□

Lemma B.5. *Suppose $n, p \to \infty$ with $\frac{p}{n} \to \gamma$, and let $\varepsilon > 0$. Then there exist $c := c_{a,\nu,\gamma} > 0$ and $N := N_{a,\nu,\gamma,\varepsilon} > 0$ and a set $G := G_{n,p} \subset \mathbb{R}^{p \times n}$ with $\mathbb{P}[Z \in G] \ge 1 - 2e^{-\frac{n}{2}}$, such that for all $t > \varepsilon$ and $n > N$,*

$$\mathbb{P}[\|M_{n,p}\| \ge \|\mu_{a,\nu,\gamma}\| + t \text{ and } Z \in G] \le e^{-cnt^2}.$$

Proof. Recall $M_{n,p} = \sqrt{\frac{\gamma(\nu - a^2)}{p}} W + \frac{a}{n}(ZZ^T - \operatorname{diag}(\|Z_i\|_2^2))$. Denote

$$\mathcal{W} = \big((w_{ii})_{1 \le i \le p}, (\sqrt{2}\operatorname{Re} w_{ij}, \sqrt{2}\operatorname{Im} w_{ij})_{1 \le i < j \le p}\big) \in \mathbb{R}^{p^2},$$

so that the entries of $\mathcal{W}$ and Z are iid $\mathcal{N}(0,1)$. Let $f : \mathbb{R}^{p^2 + np} \to \mathbb{R}$ and $f_v : \mathbb{R}^{p^2 + np} \to \mathbb{R}$ for $v \in \mathbb{C}^p$ be given by $f(\mathcal{W}, Z) = \|M_{n,p}\|$ and $f_v(\mathcal{W}, Z) = v^* M_{n,p} v$, so that $f(\mathcal{W}, Z) = \sup_{v \in \mathbb{C}^p : \|v\|_2 = 1} |f_v(\mathcal{W}, Z)|$. Note that

$$\begin{aligned}
f_v(\mathcal{W}, Z) &= \sqrt{\frac{\gamma(\nu - a^2)}{p}} \left(\sum_{i=1}^p w_{ii} |v_i|^2 + \sum_{1 \le i < j \le p} w_{ij} \overline{v_i} v_j + \overline{w_{ij}} v_i \overline{v_j} \right) + \frac{a}{n} \left(\sum_{1 \le i < j \le p} (\overline{v_i} v_j + v_i \overline{v_j}) Z_i^T Z_j \right) \\
&= \sqrt{\frac{\gamma(\nu - a^2)}{p}} \left(\sum_{i=1}^p w_{ii} |v_i|^2 + \sum_{1 \le i < j \le p} (\operatorname{Re} w_{ij})(\overline{v_i} v_j + v_i \overline{v_j}) + i(\operatorname{Im} w_{ij})(\overline{v_i} v_j - v_i \overline{v_j}) \right) \\
&\qquad + \frac{a}{n} \left(\sum_{1 \le i < j \le p} (\overline{v_i} v_j + v_i \overline{v_j}) Z_i^T Z_j \right), \\
\frac{\partial f_v(\mathcal{W}, Z)}{\partial w_{ii}} &= \sqrt{\frac{\gamma(\nu - a^2)}{p}} |v_i|^2, \quad \frac{\partial f_v(\mathcal{W}, Z)}{\partial (\sqrt{2} \operatorname{Re} w_{ij})} = \sqrt{\frac{2\gamma(\nu - a^2)}{p}} \operatorname{Re}(\overline{v_i} v_j), \\
\frac{\partial f_v(\mathcal{W}, Z)}{\partial (\sqrt{2} \operatorname{Im} w_{ij})} &= -\sqrt{\frac{2\gamma(\nu - a^2)}{p}} \operatorname{Im}(\overline{v_i} v_j), \quad \nabla_{Z_i} f_v(\mathcal{W}, Z) = \frac{2a}{n} \sum_{\substack{j=1 \\ j \ne i}}^p \operatorname{Re}(\overline{v_i} v_j) Z_j.
\end{aligned}$$

Then, for any $v \in \mathbb{C}^p$ such that $\|v\|_2 = 1$,

$$\begin{aligned}
\|\nabla f_v(\mathcal{W}, Z)\|_2^2 &= \frac{\gamma(\nu - a^2)}{p} \left(\sum_{i=1}^p |v_i|^4 + 2 \sum_{1 \le i < j \le p} |\overline{v_i} v_j|^2 \right) + \frac{4a^2}{n^2} \sum_{i=1}^p \left\| \sum_{\substack{j=1 \\ j \ne i}}^p \operatorname{Re}(\overline{v_i} v_j) Z_j \right\|_2^2 \\
&\le \frac{\gamma(\nu - a^2)}{p} \left(\sum_{i=1}^p |v_i|^2 \right)^2 + \frac{4a^2}{n^2} \sum_{i=1}^p |v_i|^2 \|Z\|^2 \|v\|_2^2
\end{aligned}$$

$$= \frac{\gamma(\nu - a^2)}{p} + \frac{4a^2\|Z\|^2}{n^2}.$$

Take $G = \{Z \in \mathbb{R}^{p\times n} : \|Z\| \le 2\sqrt{n} + \sqrt{p}\}$. Then by Corollary 5.35 of [31], $\mathbb{P}[Z \notin G] \le 2e^{-\frac{n}{2}}$. As $\mathbb{R}^{p^2} \times G$ is convex, the above inequality implies $f_v(\mathcal{W}, Z)$ is L-Lipschitz on $\mathbb{R}^{p^2} \times G$ for $L := \left(\frac{\gamma(\nu-a^2)}{p} + \frac{4a^2(2\sqrt{n}+\sqrt{p})^2}{n^2}\right)^{1/2} = O(n^{-1/2})$. Then

$$\begin{aligned} f(\mathcal{W}, Z) - f(\mathcal{W}', Z') &\le \sup_{v\in\mathbb{C}^p:\|v\|_2=1} \left(|f_v(\mathcal{W}, Z)| - |f_v(\mathcal{W}', Z')|\right) \\ &\le \sup_{v\in\mathbb{C}^p:\|v\|_2=1} \left|f_v(\mathcal{W}, Z) - f_v(\mathcal{W}', Z')\right| \le L\|(\mathcal{W}, Z) - (\mathcal{W}', Z')\|_2 \end{aligned}$$

for all $\mathcal{W}, \mathcal{W}' \in \mathbb{R}^{p^2}$ and $Z, Z' \in G$, so f is also L-Lipschitz on $\mathbb{R}^{p^2} \times G$.

Let $\tilde{f} : \mathbb{R}^{p^2+np} \to \mathbb{R}$ be the L-Lipschitz extension of f on $\mathbb{R}^{p^2} \times G$ given by Lemma B.4. Note that

$$\begin{aligned} |\mathbb{E}f(\mathcal{W}, Z) - \mathbb{E}\tilde{f}(\mathcal{W}, Z)| &= |\mathbb{E}[(f(\mathcal{W}, Z) - \tilde{f}(\mathcal{W}, Z))\mathbb{1}\{Z \notin G\}]| \\ &\le \mathbb{E}|f(\mathcal{W}, Z)\mathbb{1}\{Z \notin G\}| + \mathbb{E}|\tilde{f}(\mathcal{W}, Z)\mathbb{1}\{Z \notin G\}|. \end{aligned}$$

Lemma B.2 (with $l = 1$) implies $\mathbb{E}|f(\mathcal{W}, Z)\mathbb{1}\{Z \notin G\}| = \mathbb{E}[\|M_{n,p}\|\mathbb{1}\{Z \notin G\}] = o(1)$. As $\tilde{f}$ is L-Lipschitz,

$$|\tilde{f}(\mathcal{W}, Z)| \le |\tilde{f}(0, 0)| + L\|(\mathcal{W}, Z)\|_2 = |f(0, 0)| + L\|(\mathcal{W}, Z)\|_2 = L\|(\mathcal{W}, Z)\|_2.$$

Let $\mathcal{A}_n = \left\{\|(\mathcal{W}, Z)\|_2 \le \sqrt{2(p^2 + np)}\right\}$. As $\|(\mathcal{W}, Z)\|_2^2$ is chi-squared distributed with $p^2 + np$ degrees of freedom, a standard tail bound gives $\mathbb{P}\left[\|(\mathcal{W}, Z)\|_2^2 \ge p^2 + np + t\right] \le e^{-\frac{t^2}{8(p^2+np)}}$. Then

$$\begin{aligned} \mathbb{E}\left[\|(\mathcal{W}, Z)\|_2^2\, \mathbb{1}\{\mathcal{A}_n^C\}\right] &= \int_{p^2+np}^{\infty} \mathbb{P}\left[\|(\mathcal{W}, Z)\|_2^2 \ge p^2 + np + t\right] dt \le \int_{p^2+np}^{\infty} e^{-\frac{t^2}{8(p^2+np)}} dt \\ &= 2\sqrt{p^2 + np} \int_{\frac{\sqrt{p^2+np}}{2}}^{\infty} e^{-\frac{s^2}{2}} ds \sim 4e^{-\frac{p^2+np}{8}}. \end{aligned}$$

This implies

$$\begin{aligned} \mathbb{E}|\tilde{f}(\mathcal{W}, Z)\mathbb{1}\{Z \notin G\}| &\le \mathbb{E}[|\tilde{f}(\mathcal{W}, Z)|\mathbb{1}\{Z \notin G\}\mathbb{1}\{\mathcal{A}_n\}] + \mathbb{E}[|\tilde{f}(\mathcal{W}, Z)|\mathbb{1}\{Z \notin G\}\mathbb{1}\{\mathcal{A}_n^C\}] \\ &\le L\sqrt{2(p^2 + np)}\mathbb{P}[Z \notin G] + L\mathbb{E}\left[\|(\mathcal{W}, Z)\|_2^2\, \mathbb{1}\{\mathcal{A}_n^C\}\right]^{1/2} = o(1). \end{aligned}$$

Then $|\mathbb{E}f(\mathcal{W}, Z) - \mathbb{E}\tilde{f}(\mathcal{W}, Z)| = o(1)$, so Lemmas B.3 and B.4 imply, for all $t > \varepsilon$ and all sufficiently large n (i.e. $n > N_{a,\nu,\gamma,\varepsilon}$ independent of t),

$$\begin{aligned} &\mathbb{P}[\|M_{n,p}\| \ge \|\mu_{a,\nu,\gamma}\| + t \text{ and } Z \in G] \\ &\quad\le \mathbb{P}\left[\|M_{n,p}\| - \mathbb{E}\|M_{n,p}\| \ge t - \tfrac{\varepsilon}{2} + |\mathbb{E}f(\mathcal{W}, Z) - \mathbb{E}\tilde{f}(\mathcal{W}, Z)| \text{ and } Z \in G\right] \\ &\quad\le e^{-\frac{(t-\varepsilon/2)^2}{2L^2}} \le e^{-\frac{t^2}{8L^2}}. \end{aligned}$$

The result follows upon noting that $L = O(n^{-1/2})$. □

Proof of Proposition 4.9. Let $c > 0$ and $G \subset \mathbb{R}^{p\times n}$ be as given by Lemma B.5. Then, for any $\varepsilon > 0$,

$$\begin{aligned} \mathbb{E}[\|M_{n,p}\|^l \mathbb{1}\{Z \in G\}] &\le (\|\mu_{a,\nu,\gamma}\| + \varepsilon)^l + \mathbb{E}\left[\|M_{n,p}\|^l \mathbb{1}\{\|M_{n,p}\| \ge \|\mu_{a,\nu,\gamma}\| + \varepsilon\}\mathbb{1}\{Z \in G\}\right] \\ &= (\|\mu_{a,\nu,\gamma}\| + \varepsilon)^l + \int_{(\|\mu_{a,\nu,\gamma}\|+\varepsilon)^l}^{\infty} \mathbb{P}\left[\|M_{n,p}\|^l \ge t \text{ and } Z \in G\right] dt \end{aligned}$$

$$= (\|\mu_{a,\nu,\gamma}\| + \varepsilon)^l + \int_{\|\mu_{a,\nu,\gamma}\|+\varepsilon}^{\infty} \mathbb{P}[\|M_{n,p}\| \geq s \text{ and } Z \in G] \cdot ls^{l-1} ds$$
$$\leq (\|\mu_{a,\nu,\gamma}\| + \varepsilon)^l + l \int_{\varepsilon}^{\infty} e^{-cns^2} (\|\mu_{a,\nu,\gamma}\| + s)^{l-1} ds$$

for all sufficiently large n, where we have applied Lemma B.5. Note that

$$\begin{aligned} l \int_{\varepsilon}^{\infty} e^{-cns^2} (\|\mu_{a,\nu,\gamma}\| + s)^{l-1} ds &\leq l \int_{\varepsilon}^{\infty} e^{-cns^2 + l(\|\mu_{a,\nu,\gamma}\|+s)} ds \\ &= l e^{l\|\mu_{a,\nu,\gamma}\| + \frac{l^2}{4cn}} \int_{\varepsilon}^{\infty} e^{-cn\left(s - \frac{l}{2cn}\right)^2} ds \\ &= \frac{l e^{l\|\mu_{a,\nu,\gamma}\| + \frac{l^2}{4cn}}}{\sqrt{2cn}} \int_{\sqrt{2cn}\left(\varepsilon - \frac{l}{2cn}\right)}^{\infty} e^{-\frac{t^2}{2}} dt \\ &\sim \frac{l e^{l\|\mu_{a,\nu,\gamma}\| + \frac{l^2}{4cn}}}{2cn\left(\varepsilon - \frac{l}{2cn}\right)} e^{-cn\left(\varepsilon - \frac{l}{2cn}\right)^2} \to 0 \end{aligned}$$

for $l = O(\log n)$, so $\mathbb{E}[\|M_{n,p}\|^l \mathbb{1}\{Z \in G\}] \leq (\|\mu_{a,\nu,\gamma}\| + \varepsilon)^l + o(1)$. On the other hand, $\mathbb{P}[Z \notin G] \leq 2e^{-\frac{n}{2}}$ by Lemma B.5, so Lemma B.2 implies $\mathbb{E}[\|M_{n,p}\|^l \mathbb{1}\{Z \notin G\}] = o(1)$ for $l = O(\log n)$. Hence $\mathbb{E}[\|M_{n,p}\|^l] \leq (\|\mu_{a,\nu,\gamma}\| + \varepsilon)^l + o(1)$. As $\varepsilon > 0$ was arbitrary, this proves the desired result. □

References

[1] Z D Bai and Y Q Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability*, 21(3):1275–1294, 1993.

[2] Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*. Springer, 2010.

[3] Hari Bercovici and Dan Voiculescu. Free convolution of measures with unbounded supports. *Indiana University Mathematics Journal*, 42(3):733–773, 1993.

[4] Peter J Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

[5] T Tony Cai and Harrison H Zhou. Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40(5):2389–2420, 2012.

[6] Mireille Capitaine, Catherine Donati-Martin, Delphine Féral, and Maxime Février. Free convolution with a semi-circular distribution and eigenvalues of spiked deformations of Wigner matrices. *Electronic Journal of Probability*, 16(64):1750–1792, 2011.

[7] Mireille Capitaine and Sandrine Péché. Fluctuations at the edges of the spectrum of the full rank deformed GUE. *Probability Theory and Related Fields*, pages 1–45, 2015.

[8] Lennart Carleson. On Bernstein's approximation problem. *Proceedings of the American Mathematical Society*, 2(6):953–961, 1951.

[9] Xiuyuan Cheng and Amit Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 2(4), 2013.

[10] Yash Deshpande and Andrea Montanari. Sparse PCA via covariance thresholding. *arXiv preprint arXiv:1311.5179*, 2013.

[11] Yen Do and Van Vu. The spectrum of random kernel matrices: Universality results for rough and varying kernels. *Random Matrices: Theory and Applications*, 2(3), 2013.

[12] Ken Dykema. On certain free product factors via an extended matrix model. *Journal of Functional Analysis*, 112(1):31–60, 1993.

[13] Noureddine El Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, 36(6):2717–2756, 2008.

[14] Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, 2010.

[15] Zoltán Füredi and János Komlós. The eigenvalues of random symmetric matrices. *Combinatorica*, 1(3):233–241, 1981.

[16] Stuart Geman. A limit theorem for the norm of random matrices. *The Annals of Probability*, 8(2):252–261, 1980.

[17] Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

[18] Robert Krauthgamer, Boaz Nadler, and Dan Vilenchik. Do semidefinite relaxations really solve sparse PCA? *arXiv preprint arXiv:1306.3690*, 2013.
[19] Rafal Latala. Some estimates of norms of random matrices. *Proceedings of the American Mathematical Society*, 133(5):1273–1282, 2005.
[20] Ji Oon Lee and Kevin Schnelli. Edge universality for deformed Wigner matrices. *arXiv preprint arXiv:1407.8015*, 2014.
[21] Hans Maassen. Addition of freely independent random variables. *Journal of Functional Analysis*, 106(2):409–438, 1992.
[22] Camille Male. The norm of polynomials in large random and deterministic matrices. *Probability Theory and Related Fields*, 154(3-4):477–532, 2012.
[23] Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Sbornik: Mathematics*, 1(4):457–483, 1967.
[24] Natesh S Pillai and Jun Yin. Universality of covariance matrices. *The Annals of Applied Probability*, 24(3):935–1001, 2014.
[25] Adam J Rothman, Elizaveta Levina, and Ji Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.
[26] Jun Shao, Yazhen Wang, Xinwei Deng, and Sijian Wang. Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics*, 39(2):1241–1265, 2011.
[27] Tatyana Shcherbina. On universality of local edge regime for the deformed Gaussian unitary ensemble. *Journal of Statistical Physics*, 143(3):455–481, 2011.
[28] Gabor Szegö. *Orthogonal Polynomials*. American Mathematical Society, 1939.
[29] Terence Tao. *Topics in Random Matrix Theory*. American Mathematical Society, 2012.
[30] Craig A Tracy and Harold Widom. On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics*, 177(3):727–754, 1996.
[31] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, pages 210–268. Cambridge University Press, 2012.
[32] Dan Voiculescu. Addition of certain non-commuting random variables. *Journal of Functional Analysis*, 66(3):323–346, 1986.
[33] Dan Voiculescu. Limit laws for random matrices and free products. *Inventiones mathematicae*, 104(1):201–220, 1991.
[34] Y Q Yin, Z D Bai, and P R Krishnaiah. On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probability Theory and Related Fields*, 78(4):509–521, 1988.