Inner-product Kernels are Asymptotically Equivalent to Binary Discrete Kernels

Zhenyu Liao[1,2,*] **Romain Couillet**[2,1]
[1] Laboratoire des Signaux et Système, CentraleSupélec, University Paris-Saclay, France.
[2] G-STATS Data Science Chair, GIPSA-lab, University Grenobles-Alpes, France.

Abstract

This article investigates the eigenspectrum of the inner product-type kernel matrix $\sqrt{p}\mathbf{K} = \{f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})\}_{i,j=1}^n$ under a binary mixture model in the high dimensional regime where the number of data n and their dimension p are both large and comparable. Based on recent advances in random matrix theory, we show that, for a wide range of nonlinear functions f, the eigenspectrum behavior is asymptotically equivalent to that of an (at most) cubic function. This sheds new light on the understanding of nonlinearity in large dimensional problems. As a byproduct, we propose a simple function prototype valued in $(-1, 0, 1)$ that, while reducing both storage memory and running time, achieves the same (asymptotic) classification performance as any arbitrary function f.

1 Introduction

Multivariate mixture models, especially Gaussian mixtures, play a fundamental role in statistics and have received significant research attention in the machine learning community [KMV16, ABDH+18], mainly due to the convenient statistical properties of Gaussian and sub-Gaussian distributions. More generic mixture models, however, are somehow less covered.

On the other hand, in the study of large random matrices, one is able to reach in some cases "universal" results in the sense that the (asymptotic) statistical behavior of the object of interest is *independent* of the underlying distribution of the random matrix. Intuitively speaking, the "squared" number of degrees of freedom in large matrices (e.g., sample covariance matrices $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} \in \mathbb{R}^{p\times p}$ based on n observations of dimension p arranged in the columns of $\mathbf{X}$) and their *independent* interactions induce fast versions of central limit theorems irrespective of the data distribution, resulting in universal statistical results.

In this paper, we consider the eigenspectrum behavior of the inner product "properly scaled" (see details below) kernel matrix $\mathbf{K}_{ij} = f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})/\sqrt{p}$, for n data $\mathbf{x}_i \in \mathbb{R}^p$ arising from an affine transformation of i.i.d. random vectors with zero mean, unit covariance and bounded higher order moments. Under this setting and some mild regularity condition for the nonlinear function f, the spectrum of $\mathbf{K}$ can be shown to only depend on *three* parameters of f, in the regime where n, p are both large and comparable.

The theoretical study of the eigenspectrum of large random matrices serves as a basis to understand many practical statistical learning algorithms, among which are kernel spectral clustering [NJW02] or sparse principle component analysis (PCA) [JL09]. In the large n, p regime, various types of "random kernel matrices" have been studied from a spectral random matrix viewpoint. In [EK10b] the authors considered kernel matrices based on the Euclidean distance $f(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)$ or the inner product $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p)$ between data vectors, and study the eigenvalue distribution by essentially "linearizing"

[*] Work done during a research visit at GIPSA lab, University of Grenoble Alpes, France.

Preprint. Under review.

the nonlinear function f via a Taylor expansion, which naturally demands f to be locally smooth. Later in [CBG16] the authors followed the same technical approach and considered a more involved Gaussian mixture model for the $\mathbf{x}_i$, providing rich insights into the impact of nonlinearity in kernel spectral clustering application.

Nonetheless, these results are in essence built upon a local expansion of the nonlinear function f, which follows from the "concentration" of the similarity measures $\|\mathbf{x}_i - \mathbf{x}_j\|^2/p$ or $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p$ around a *single* value of the smooth domain of f, therefore disregarding most of the domain of f. In this article, following [CS13, DV13], we study the inner product kernel $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})/\sqrt{p}$ which avoids the concentration effects with the more natural $\sqrt{p}$ normalization. With the flexible tool of orthogonal polynomials, we are able to prove universal results which solely depend on the first two order moments of the data distribution and allow for nonlinear functions f that need not even be differentiable. As a practical outcome of our theoretical results, we propose an extremely simple piecewise constant function which is spectrally equivalent and thus performs equally well as arbitrarily complex functions f, while inducing enormous gains in both storage and computational complexity.

The remainder of this article is organized as follows. In Section 2 we introduce the object of interest together with necessary assumptions to work along with. Our main theoretical findings are presented in Section 3 with intuitive ideas, while detailed proofs are deferred to the supplementary material due to space limitation. In Section 4 we discuss the practical consequence of our theoretical findings and propose our piecewise constant function prototype which works in a universal manner for kernel spectrum-based applications, for the system model under consideration. The article closes with concluding remarks and envisioned extensions in Section 5.

Notations: Boldface lowercase (uppercase) characters stand for vectors (matrices). The notation $(\cdot)^\mathsf{T}$ denotes the transpose operator. The norm $\|\cdot\|$ is the Euclidean norm for vectors and the operator norm for matrices, and we denote $\|\mathbf{A}\|_\infty = \max_{i,j}|\mathbf{A}_{ij}|$ as well as $\|\cdot\|_F$ the Frobenius norm: $\|\mathbf{A}\|_F^2 = \operatorname{tr}(\mathbf{A}\mathbf{A}^\mathsf{T})$. ξ is often used to denote standard Gaussian random variable, i.e., $\xi \sim \mathcal{N}(0,1)$.

2 System model and preliminaries

2.1 Basic setting

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be n feature vectors drawn independently from the following two-class ($\mathcal{C}_1$ and $\mathcal{C}_2$) mixture model:

$$\begin{cases} \mathcal{C}_1: & \mathbf{x} = \boldsymbol{\mu}_1 + (\mathbf{I}_p + \mathbf{E}_1)^{\frac{1}{2}}\mathbf{z} \\ \mathcal{C}_2: & \mathbf{x} = \boldsymbol{\mu}_2 + (\mathbf{I}_p + \mathbf{E}_2)^{\frac{1}{2}}\mathbf{z} \end{cases} \tag{1}$$

each having cardinality $n/2$,[2] for some deterministic $\boldsymbol{\mu}_a \in \mathbb{R}^p$, $\mathbf{E}_a \in \mathbb{R}^{p\times p}$, $a = 1, 2$ and random vector $\mathbf{z} \in \mathbb{R}^p$ having i.i.d. entries of zero mean, unit variance and bounded moments. To ensure that the information of $\boldsymbol{\mu}_a, \mathbf{E}_a$ is neither (asymptotically) too simple nor impossible to be extracted from the noisy features[3], we work (as in [CLM18]) under the following assumption.

Assumption 1 (Non-trivial classification)**.** *As $n \to \infty$, we have for $a \in \{1, 2\}$*

(i) $p/n = c \to \bar{c} \in (0, \infty)$,

(ii) $\|\boldsymbol{\mu}_a\| = O(1)$, $\|\mathbf{E}_a\| = O(p^{-1/4})$, $|\operatorname{tr}(\mathbf{E}_a)| = O(\sqrt{p})$ *and* $\|\mathbf{E}_a\|_F^2 = O(\sqrt{p})$.

Following [EK10a, CS13] we consider the following nonlinear random inner-product matrix

$$\mathbf{K} = \left\{\delta_{i\neq j} f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})/\sqrt{p}\right\}_{i,j=1}^n \tag{2}$$

for function $f : \mathbb{R} \mapsto \mathbb{R}$ satisfying some regularity conditions (detailed later in Assumption 2). As in [EK10a, CS13], the diagonal elements $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_i/\sqrt{p})$ have been discarded. Indeed, under

[2] We restrict ourselves to binary classification for readability, but the presented results easily extend to a multi-class setting.

[3] We refer the readers to Section A in Supplementary Material for a more detailed discussion on this point.

Assumption 1, $\mathbf{x}_i^\mathsf{T}\mathbf{x}_i/\sqrt{p} = O(\sqrt{p})$ which is an "improper scaling" for the evaluation by f (unlike $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p}$ which properly scales as $O(1)$ for all $i \neq j$).

In the absence of discriminative information (null model), i.e., if $\boldsymbol{\mu}_a = \mathbf{0}$ and $\mathbf{E}_a = \mathbf{0}$ for $a = 1, 2$, we write $\mathbf{K} = \mathbf{K}_N$ with

$$[\mathbf{K}_N]_{ij} = \delta_{i\neq j} f(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})/\sqrt{p}. \tag{3}$$

Letting $\xi_p \equiv \mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p}$, by the central limit theorem, $\xi_p \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ as $p \to \infty$. As such, the $[\mathbf{K}_N]_{ij}$, $1 \leq i \neq j \leq n$, asymptotically behave like a family of *dependent* standard Gaussian variables to which f is applied. In order to analyze the joint behavior of this family, we shall exploit some useful concepts of the theory of orthogonal polynomials and, in particular, of the class of Hermite polynomials defined with respect to the standard Gaussian distribution [AS65, AAR00].

2.2 The Orthogonal Polynomial Framework

For a real probability measure μ, we denote the set of orthogonal polynomials with respect to the scalar product $\langle f, g\rangle = \int fg d\mu$ as $\{P_l(x), l = 0, 1, \ldots\}$, obtained from the Gram-Schmidt procedure on the monomials $\{1, x, x^2, \ldots\}$ such that $P_0(x) = 1$, P_l is of degree l and $\langle P_{l_1}, P_{l_2}\rangle = \delta_{l_1 - l_2}$. By the Riesz-Fisher theorem [Rud64, Theorem 11.43], for any function $f \in L^2(\mu)$, the set of squared integrable functions with respect to $\langle \cdot, \cdot \rangle$, one can formally expand f as

$$f(x) \sim \sum_{l=0}^{\infty} a_l P_l(x), \quad a_l = \int_{\mathbb{R}} f(x) P_l(x) d\mu(x) \tag{4}$$

where "$f \sim \sum_{l=0}^{\infty} P_l$" indicates that $\|f - \sum_{l=0}^{N} P_l\| \to 0$ as $N \to \infty$ (and $\|f\|^2 = \langle f, f\rangle$).

To investigate the asymptotic behavior of $\mathbf{K}$ and $\mathbf{K}_N$ as $n, p \to \infty$, we position ourselves under the following technical assumption.

Assumption 2. *For each* p*, let* $\xi_p = \mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p}$ *and let* $\{P_{l,p}(x),\ l \geq 0\}$ *be the set of orthonormal polynomials with respect to the probability measure* μ_p *of* ξ_p.[4] *For* $f \in L^2(\mu_p)$ *for each* p*, i.e.,*

$$f(x) \sim \sum_{l=0}^{\infty} a_{l,p} P_{l,p}(x)$$

for $a_{l,p}$ *defined in* (4)*, we demand that*

- *(i)* $\sum_{l=0}^{\infty} a_{l,p} P_{l,p}(x)\mu_p(dx)$ *converges in* $L^2(\mu_p)$ *to* $f(x)$ *uniformly over large* p*, i.e., for any* $\epsilon > 0$ *there exists* L *such that for all* p *large,*

$$\left\| f - \sum_{l=0}^{L} a_{l,p} P_{l,p} \right\|_{L^2(\mu_p)}^2 = \sum_{l=L+1}^{\infty} |a_{l,p}|^2 \leq \epsilon,$$

- *(ii) as* $p \to \infty$, $\sum_{l=1}^{\infty} |a_{l,p}|^2 \to \nu \in [0, \infty)$*. Moreover, for* $l = 0, 1, 2$, $a_{l,p}$ *converges and we denote* a_0, a_1 *and* a_2 *their limits, respectively.*
- *(iii)* $a_0 = 0$.

Since $\xi_p \to \mathcal{N}(0,1)$, the limiting parameters a_0, a_1, a_2 and ν are simply (generalized) moments of the standard Gaussian measure involving f. Precisely,

$$a_0 = \mathbb{E}[f(\xi)],\ a_1 = \mathbb{E}[\xi f(\xi)],\ a_2 = \frac{\mathbb{E}[(\xi^2 - 1)f(\xi)]}{\sqrt{2}} = \frac{\mathbb{E}[\xi^2 f(\xi)] - a_0}{\sqrt{2}},\ \nu = \mathrm{Var}[f(\xi)] \geq a_1^2 + a_2^2$$

for $\xi \sim \mathcal{N}(0,1)$. These parameters are of crucial significance in determining the eigenspectrum behavior of $\mathbf{K}$. Note that a_0 will not affect the classification performance, as described below.

Remark 1 (On a_0). *In the present case of balanced mixtures (equal cardinalities for* $\mathcal{C}_1$ *and* $\mathcal{C}_2$*),* a_0 *contributes to the polynomial expansion of* $\mathbf{K}_N$ *(and* $\mathbf{K}$*) as a non-informative rank-1 perturbation of the form* $a_0(\mathbf{1}_n\mathbf{1}_n^\mathsf{T} - \mathbf{I}_n)/\sqrt{p}$*. Since* $\mathbf{1}_n$ *is orthogonal to the "class-information vector"* $[\mathbf{1}_{n/2}, -\mathbf{1}_{n/2}]$*, its presence does not impact the classification performance.*[5]

[4] Note that μ_p is merely standard Gaussian in the large p limit.

[5] If mixtures are unbalanced, the vector $\mathbf{1}_n$ may tend to "pull" eigenvectors aligned to $[\mathbf{1}_{n_1}, -\mathbf{1}_{n_2}]$, with n_i the cardinality in $\mathcal{C}_i$, so away from purely noisy eigenvectors and thereby impacting classification performance. See [CBG16] for similar considerations.

2.3 Limiting spectrum of $\mathbf{K}_N$

It was shown in [CS13, DV13] that, for independent $\mathbf{z}_i$'s with independent entries, the *empirical spectral measure* $\mathcal{L}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\lambda_i(\mathbf{K}_N)}$ of the null model $\mathbf{K}_N$ has an asymptotically deterministic behavior as $n, p \to \infty$ with $p/n \to \bar{c} \in (0, \infty)$.

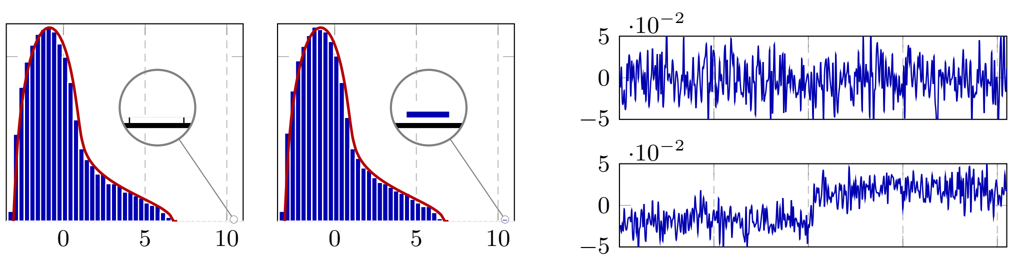
Theorem 1 ([CS13, DV13])**.** *Let $p/n = c \to \bar{c} \in (0,\infty)$ and Assumption 2 hold. Then, the empirical spectral measure $\mathcal{L}_n$ of $\mathbf{K}_N$ defined in* (3) *converges weakly and almost surely to a probability measure $\mathcal{L}$. The latter is uniquely defined through its Stieltjes transform $m : \mathbb{C}^+ \to \mathbb{C}^+$, $z \mapsto \int (t-z)^{-1}\mathcal{L}(dt)$, given as the unique solution in $\mathbb{C}^+$ of the (cubic) equation*[6]

$$-\frac{1}{m(z)} = z + \frac{a_1^2 m(z)}{c + a_1 m(z)} + \frac{\nu - a_1^2}{c} m(z).$$

Theorem 1 is "universal" with respect to the law of the (independent) entries of $\mathbf{z}_i$. While universality is classical in random matrix results, with mostly first and second order statistics involved, the present universality result is much less obvious since (i) the nonlinear application $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})$ depends in an intricate manner on all moments of $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$ and (ii) the entries of $\mathbf{K}_N$ are strongly dependent. In essence, universality still holds here because the convergence speed to Gaussian of $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p}$ is sufficiently fast to compensate the residual impact of higher order moments in the spectrum of $\mathbf{K}_N$.

As an illustration, Figure 1a compares the empirical spectral measure of $\mathbf{K}_N$ to the limiting measure μ of Theorem 1.[7]

From a technical viewpoint, the objective of the article is to go beyond the null model described in Theorem 1 by providing a tractable *random matrix equivalent* $\tilde{\mathbf{K}}$ for the kernel matrix $\mathbf{K}$, in the sense that $\|\mathbf{K} - \tilde{\mathbf{K}}\| \to 0$ almost surely in operator norm, as $n, p \to \infty$. This convergence allows one to identify the eigenvalues and isolated eigenvectors (that can be used for spectral clustering purpose) of $\mathbf{K}$ by means of those of $\tilde{\mathbf{K}}$, see for instance [HJ12, Corollary 4.3.15]. More importantly, while not visible from the expression of $\mathbf{K}$, the impact of the mixture model $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \mathbf{E}_1, \mathbf{E}_2)$ on $\mathbf{K}$ is readily accessed from $\tilde{\mathbf{K}}$ and easily related to the Hermite coefficients (a_1, a_2, ν) of f. This allows us to further investigate how the choice of f impacts the asymptotically feasibility and efficiency of spectral clustering from the top eigenvectors of $\mathbf{K}$.



(a) Eigenvalues of $\mathbf{K}_N$ (b) Eigenvalues of $\mathbf{K}$ (c) Top eigenvectors of $\mathbf{K}_N$ (top) and $\mathbf{K}$ (bottom)

Figure 1: Eigenvalue distribution and top eigenvector of $\mathbf{K}_N$ and $\mathbf{K}$, together with the limiting spectral measure $\mathcal{L}$ (from Theorem 1) in red; $f(x) = \operatorname{sign}(x)$, Gaussian $\mathbf{z}_i$, $n = 2\,048$, $p = 512$, $\boldsymbol{\mu}_1 = -[3/2;\ \mathbf{0}_{p-1}] = -\boldsymbol{\mu}_2$ and $\mathbf{E}_1 = \mathbf{E}_2 = \mathbf{0}$. $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$ with $\mathbf{x}_1, \ldots, \mathbf{x}_{n/2} \in \mathcal{C}_1$ and $\mathbf{x}_{n/2+1}, \ldots, \mathbf{x}_n \in \mathcal{C}_2$.

3 Theoretical results

The main idea for the asymptotic analysis of $\mathbf{K}$ comes in two steps: (i) first, by an expansion of $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$ as a function of $\mathbf{z}_i, \mathbf{z}_j$ and the statistical mixture model parameters $\boldsymbol{\mu}, \mathbf{E}$, we decompose $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$ (under

[6] $\mathbb{C}^+ \equiv \{z \in \mathbb{C},\ \Im[z] > 0\}$. We also recall that, for $m(z)$ the Stieltjes transform of a measure μ, μ can be obtained from $m(z)$ via $\mu([a,b]) = \lim_{\epsilon\downarrow 0} \frac{1}{\pi}\int_a^b \Im[m(x + \imath\epsilon)]dx$ for all $a < b$ continuity points of μ.

[7] For all figures in this article, the eigenvalues (that produce the empirical histograms) as well as the associated eigenvectors are computed by MATLAB's `eig(s)` function and correspond to a *single* realization of the random kernel matrix $\mathbf{K}$ or $\mathbf{K}_N$.

Assumption 1) into successive orders of magnitudes with respect to p; this, as we will show, further allows for a Taylor expansion of $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})$ for at least twice differentiable functions f around its dominant term $f(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})$. Then, (ii) we rely on the orthogonal polynomial approach of [CS13] to "linearize" the resulting matrix terms $\{f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})\}$, $\{f'(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})\}$ and $\{f''(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})\}$ (all terms corresponding to higher order derivatives asymptotically vanish) and use Assumption 2 to extend the result to arbitrary square-summable f.

Our main conclusion is that $\mathbf{K}$ asymptotically behaves like a matrix $\tilde{\mathbf{K}}$ following a so-called "spiked random matrix model" in the sense that $\tilde{\mathbf{K}} = \mathbf{K}_N + \tilde{\mathbf{K}}_I$ is the sum of the full-rank "noise" matrix $\mathbf{K}_N$ having compact limiting spectrum (the support of $\mathcal{L}$) and a low-rank "information" matrix $\tilde{\mathbf{K}}_I$ [BAP05, BGN11].

3.1 Information-plus-noise decomposition of $\mathbf{K}$

We first show that $\mathbf{K}$ can be asymptotically approximated as $\mathbf{K}_N + \mathbf{K}_I$ with $\mathbf{K}_N$ defined in (3) and $\mathbf{K}_I$ an additional (so far full-rank) term containing the statistical information of the mixture model.

As announced, we start by decomposing $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$ into a sequence of terms of successive orders of magnitude using Assumption 1 and $\mathbf{x}_i = \boldsymbol{\mu}_a + (\mathbf{I}_p + \mathbf{E}_a)^{\frac{1}{2}}\mathbf{z}_i$, $\mathbf{x}_j = \boldsymbol{\mu}_b + (\mathbf{I}_p + \mathbf{E}_b)^{\frac{1}{2}}\mathbf{z}_j$ for $\mathbf{x}_i \in \mathcal{C}_a$ and $\mathbf{x}_j \in \mathcal{C}_b$. We have precisely, for $i \neq j$,

$$
\begin{aligned}
\frac{\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{\sqrt{p}} &= \frac{\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\mu}_b}{\sqrt{p}} + \frac{1}{\sqrt{p}}(\boldsymbol{\mu}_a^\mathsf{T}(\mathbf{I}_p + \mathbf{E}_b)^{\frac{1}{2}}\mathbf{z}_j + \boldsymbol{\mu}_b^\mathsf{T}(\mathbf{I}_p + \mathbf{E}_a)^{\frac{1}{2}}\mathbf{z}_i) + \frac{1}{\sqrt{p}}\mathbf{z}_i^\mathsf{T}(\mathbf{I}_p + \mathbf{E}_a)^{\frac{1}{2}}(\mathbf{I}_p + \mathbf{E}_b)^{\frac{1}{2}}\mathbf{z}_j \\
&= \underbrace{\frac{\mathbf{z}_i^\mathsf{T}\mathbf{z}_j}{\sqrt{p}}}_{O(1)} + \underbrace{\frac{\mathbf{z}_i^\mathsf{T}(\mathbf{E}_a + \mathbf{E}_b)\mathbf{z}_j}{2\sqrt{p}}}_{\equiv \mathbf{A}_{ij} = O(p^{-1/4})} + \underbrace{\frac{\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\mu}_b + \boldsymbol{\mu}_a^\mathsf{T}\mathbf{z}_j + \boldsymbol{\mu}_b^\mathsf{T}\mathbf{z}_i}{\sqrt{p}} - \frac{\mathbf{z}_i^\mathsf{T}(\mathbf{E}_a - \mathbf{E}_b)^2\mathbf{z}_j}{8\sqrt{p}}}_{\equiv \mathbf{B}_{ij} = O(p^{-1/2})} + o(p^{-1/2})
\end{aligned}
\tag{5}
$$

where in particular we performed a Taylor expansion of $(\mathbf{I}_p + \mathbf{E}_a)^{\frac{1}{2}}$ (since $\|\mathbf{E}_a\| = O(p^{-\frac{1}{4}})$) around $\mathbf{I}_p$, and used the fact that with high probability $\mathbf{z}_i^\mathsf{T}\mathbf{E}_a\mathbf{z}_j = O(p^{1/4})$ and $\mathbf{z}_i^\mathsf{T}(\mathbf{E}_a - \mathbf{E}_b)^2\mathbf{z}_j = O(1)$.

As a consequence of this expansion, for at least twice differentiable $f \in L^2(\mu_p)$, we have

$$
\mathbf{K}_{ij} = \frac{f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})}{\sqrt{p}} = \frac{f(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})}{\sqrt{p}} + \frac{f'(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})}{\sqrt{p}}(\mathbf{A}_{ij} + \mathbf{B}_{ij}) + \frac{f''(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})}{2\sqrt{p}}\mathbf{A}_{ij}^2 + o(p^{-1})
$$

where $o(p^{-1})$ is understood with high probability and uniformly over $i, j \in \{1, \ldots, n\}$. This *entry-wise* expansion up to order $o(p^{-1})$ is sufficient since, *matrix-wise*, if $\mathbf{A}_{ij} = o(p^{-1})$ uniformly on i, j, from $\|\mathbf{A}\| \leq p\|\mathbf{A}\|_\infty = p\max_{i,j}|\mathbf{A}_{ij}|$, we have $\|\mathbf{A}\| = o(1)$ as $n, p \to \infty$.

In the particular case where f is a monomial of degree $k \geq 2$, this implies the following result.

Proposition 1 (Monomial f). *Under Assumptions 1–2, let $f(x) = x^k$, $k \geq 2$. Then, as $n, p \to \infty$,*

$$
\|\mathbf{K} - (\mathbf{K}_N + \mathbf{K}_I)\| \to 0 \tag{6}
$$

almost surely, with $\mathbf{K}_N$ defined in (3) *and*

$$
\mathbf{K}_I = \frac{k}{\sqrt{p}}(\mathbf{Z}^\mathsf{T}\mathbf{Z}/\sqrt{p})^{\circ(k-1)} \circ (\mathbf{A} + \mathbf{B}) + \frac{k(k-1)}{2\sqrt{p}}(\mathbf{Z}^\mathsf{T}\mathbf{Z}/\sqrt{p})^{\circ(k-2)} \circ (\mathbf{A})^{\circ 2} \tag{7}
$$

for $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p \times n}$ and $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ defined in (5) *with $\mathbf{A}_{ii} = \mathbf{B}_{ii} = 0$. Here $\mathbf{X} \circ \mathbf{Y}$ denotes the Hadamard product between $\mathbf{X}, \mathbf{Y}$ and $\mathbf{X}^{\circ k}$ the k-th Hadamard power, i.e., $[\mathbf{X}^{\circ k}]_{ij} = (\mathbf{X}_{ij})^k$.*

Since $f \in L^2(\mu)$ can be decomposed into its Hermite polynomials, Proposition 1 along with Theorem 1 allows for an asymptotic quantification of $\mathbf{K}$. However, the expression of $\mathbf{K}_I$ in (7) does not so far allow for a thorough understanding of the spectrum of $\mathbf{K}$, due to (i) the delicate Hadamard products between purely random ($\mathbf{Z}^\mathsf{T}\mathbf{Z}$) and informative matrices ($\mathbf{A}, \mathbf{B}$) and (ii) the fact that $\mathbf{K}_I$ is full rank (so that the resulting spectral properties of $\mathbf{K}_N + \mathbf{K}_I$ remains intractable). We next show that, as $n, p \to \infty$, $\mathbf{K}_I$ admits a tractable low-rank approximation $\tilde{\mathbf{K}}_I$, thereby leading to a spiked-model approximation of $\mathbf{K}$.

3.2 Spiked-model approximation of K

Let us then consider $\mathbf{K}_I$ defined in (7), the (i, j) entry of which can be written as the sum of terms containing $\boldsymbol{\mu}_a, \boldsymbol{\mu}_b$ (treated separately) and random variables of the type

$$\phi = \frac{C}{\sqrt{p}}(\mathbf{x}^\mathsf{T}\mathbf{y}/\sqrt{p})^\alpha(\mathbf{x}^\mathsf{T}\mathbf{F}\mathbf{y})^\beta$$

for independent random vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ with i.i.d. zero mean, unit variance and finite moments (uniformly on p) entries, deterministic $\mathbf{F} \in \mathbb{R}^{p\times p}$, $C \in \mathbb{R}$, $\alpha \in \mathbb{N}$ and $\beta \in \{1, 2\}$.

For Gaussian $\mathbf{x}, \mathbf{y}$, the expectation of ϕ can be explicitly computed via an integral trick [Wil97, LLC18]. For more generic $\mathbf{x}, \mathbf{y}$ with i.i.d. bounded moment entries, a combinatorial argument controls the higher order moments of the expansion which asymptotically result in (matrix-wise) vanishing terms. See Sections B–C of the supplementary material. This leads to the following result.

Proposition 2 (Low rank asymptotics of $\mathbf{K}_I$)**.** *Under Assumptions 1–2, for $f(x) = x^k$, $k \geq 2$,*

$$\|\mathbf{K}_I - \tilde{\mathbf{K}}_I\| \to 0$$

almost surely as $n, p \to \infty$, for $\mathbf{K}_I$ defined in (7) *and*

$$\tilde{\mathbf{K}}_I = \begin{cases} \frac{k!!}{p}(\mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T} + \mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{Z} + \mathbf{Z}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T}), & \textit{for } k \textit{ odd} \\ \frac{k(k-1)!!}{2p}\mathbf{J}\left(\mathbf{T} + \mathbf{S}\right)\mathbf{J}^\mathsf{T}, & \textit{for } k \textit{ even} \end{cases} \tag{8}$$

where[8]

$$\mathbf{M} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2] \in \mathbb{R}^{p\times 2},\ \mathbf{T} = \{\mathrm{tr}(\mathbf{E}_a + \mathbf{E}_b)/\sqrt{p}\}_{a,b=1}^2,\ \mathbf{S} = \{\mathrm{tr}(\mathbf{E}_a\mathbf{E}_b)/\sqrt{p}\}_{a,b=1}^2 \in \mathbb{R}^{2\times 2}$$

and $\mathbf{J} = [\mathbf{j}_1, \mathbf{j}_2] \in \mathbb{R}^{n\times 2}$ with $\mathbf{j}_a \in \mathbb{R}^n$ the canonical vector of class $\mathcal{C}_a$, i.e., $[\mathbf{j}_a]_i = \delta_{\mathbf{x}_i\in\mathcal{C}_a}$.

We refer the readers to Section C of the supplementary material for a detailed exposition of the proof.

Proposition 2 states that $\mathbf{K}_I$ is asymptotically equivalent to $\tilde{\mathbf{K}}_I$ that is of rank at most two.[9] Note that the eigenvectors of $\tilde{\mathbf{K}}_I$ are linear combinations of the vectors $\mathbf{j}_1, \mathbf{j}_2$ and thus provide the data classes.

From the expression of $\tilde{\mathbf{K}}_I$, quite surprisingly, it appears that for $f(x) = x^k$, depending on whether k is odd or even, either only the information in means ($\mathbf{M}$) or only in covariance ($\mathbf{T}$ and $\mathbf{S}$) can be (asymptotically) preserved.

By merely combining the results of Propositions 1–2, the latter can be easily extended to polynomial f. Then, by considering $f(x) = P_\kappa(x)$, the Hermite polynomial of degree κ, it can be shown that, quite surprisingly, one has $\tilde{\mathbf{K}}_I = \mathbf{0}$ if $\kappa > 2$ (see Section D of the supplementary material). As such, using the Hermite polynomial expansion $P_0, P_1, \ldots$ of an arbitrary $f \in L^2(\mu)$ satisfying Assumption 2 leads to a very simple expression of our main result.

Theorem 2 (Spiked-model approximation of $\mathbf{K}$)**.** *For an arbitrary $f \in L^2(\mu)$ with $f \sim \sum_{l=0}^\infty a_l P_l(x)$, under Assumptions 1–2,*

$$\|\mathbf{K} - \tilde{\mathbf{K}}\| \to 0, \quad \tilde{\mathbf{K}} = \mathbf{K}_N + \tilde{\mathbf{K}}_I$$

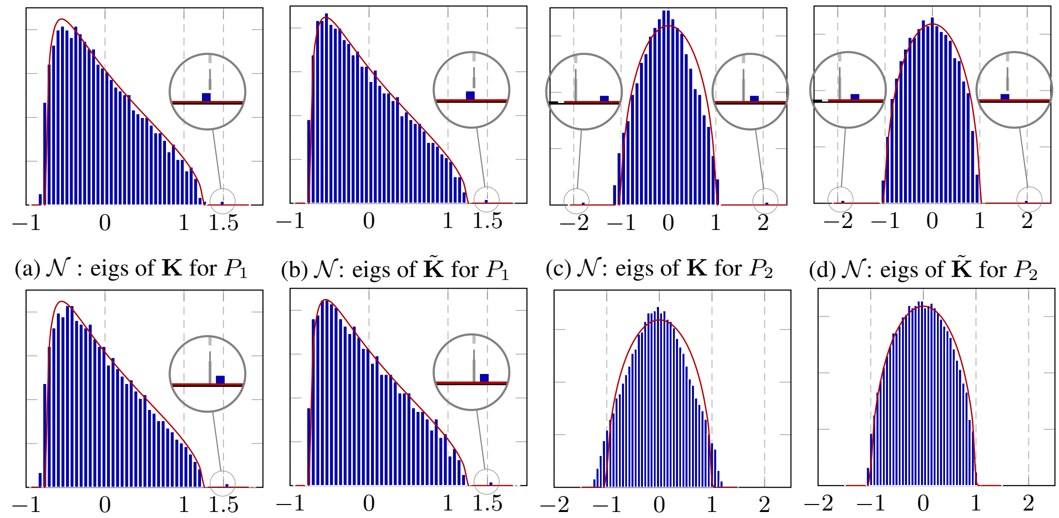
with $\mathbf{K}_N$ defined in (3) *and*

$$\tilde{\mathbf{K}}_I = \frac{a_1}{p}(\mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T} + \mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{Z} + \mathbf{Z}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T}) + \frac{a_2}{p}\mathbf{J}(\mathbf{T} + \mathbf{S})\mathbf{J}^\mathsf{T}. \tag{9}$$

The detailed proof of Theorem 2 is provided in Section D of the supplementary material.

[8]For mental reminder, $\mathbf{M}$ stands for *means*, $\mathbf{T}$ accounts for the difference in *traces* of covariance matrices and $\mathbf{S}$ for the "shapes" of the covariances.

[9]Note that, as defined, $\tilde{\mathbf{K}}_I$ has non-zero diagonal elements, while $[\mathbf{K}_I]_{ii} = 0$. This is not contradictory as the diagonal matrix $\mathrm{diag}(\tilde{\mathbf{K}}_I)$ has vanishing norm and can thus be added without altering the approximation $\|\mathbf{K}_I - \tilde{\mathbf{K}}_I\| \to 0$; it however appears convenient as it ensures that $\tilde{\mathbf{K}}_I$ is low rank (while without its diagonal, $\tilde{\mathbf{K}}_I$ is full rank).



(a) $\mathcal{N}$: eigs of $\mathbf{K}$ for P_1 (b) $\mathcal{N}$: eigs of $\tilde{\mathbf{K}}$ for P_1 (c) $\mathcal{N}$: eigs of $\mathbf{K}$ for P_2 (d) $\mathcal{N}$: eigs of $\tilde{\mathbf{K}}$ for P_2

(e) Stud: eigs of $\mathbf{K}$ for P_1 (f) Stud: eigs of $\tilde{\mathbf{K}}$ for P_1 (g) Stud: eigs of $\mathbf{K}$ for P_3 (h) Stud: eigs of $\tilde{\mathbf{K}}$ for P_3

Figure 2: Eigenvalue distributions of $\mathbf{K}$ and $\tilde{\mathbf{K}}$ from Theorem 2 (blue) and $\mathcal{L}$ from Theorem 1 (red), for $\mathbf{z}_i$ with Gaussian (top) or Student-t with degree of freedom 7 (bottom) entries; functions $f(x) = P_1(x) = x$, $f(x) = P_2(x) = (x^2 - 1)\sqrt{2}$, $f(x) = P_3(x) = (x^3 - 3x)/\sqrt{6}$; $n = 2\,048$, $p = 8\,192$, $\boldsymbol{\mu}_1 = -[2; \mathbf{0}_{p-1}] = -\boldsymbol{\mu}_2$ and $\mathbf{E}_1 = -10\mathbf{I}_p/\sqrt{p} = -\mathbf{E}_2$.

Figure 2 compares the spectra of $\mathbf{K}$ and $\tilde{\mathbf{K}}$ for random vectors with independent Gaussian or Student-t entries, for the first three (normalized) Hermite polynomials $P_1(x)$, $P_2(x)$ and $P_3(x)$. These numerical evidences validate Theorem 2: only for $P_1(x)$ and $P_2(x)$ is an isolated eigenvalue observed. Besides, as shown in the bottom display of Figure 1c, the corresponding eigenvector is, as expected, a noisy version of linear combinations of $\mathbf{j}_1, \mathbf{j}_2$.

Remark 2 (Even and odd f)**.** *While* $\operatorname{rank}(\tilde{\mathbf{K}}_I) \leq 4$ *(as the sum of two rank-two terms), in Figure 2* no more than *two isolated eigenvalues are observed (for $f = P_1$ only one on the right side, for $f = P_2$ one on each side). This follows from $a_2 = 0$ when $f = P_1$ and $a_1 = 0$ for $f = P_2$. More generally, for f* odd *($f(-x) = -f(x)$), $a_2 = 0$ and the statistical information on covariances (through* $\mathbf{E}$*) asymptotically vanishes in* $\mathbf{K}$*; for f* even *($f(-x) = f(x)$), $a_1 = 0$ and information about the means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ vanishes. Thus, only f neither odd nor even can preserve both first and second order discriminating statistics (e.g., the popular ReLU function $f(x) = \max(0, x)$). This was previously remarked in [LC18] based on a local expansion of smooth f in a similar setting.*

4 Practical consequences: universality of binary kernels

As a direct consequence of Theorem 2, the performance of spectral clustering for large dimensional mixture models of the type (1) only depends on the *three* parameters of the nonlinear function f: $a_1 = \mathbb{E}[\xi f(\xi)]$, $a_2 = \mathbb{E}[\xi^2 f(\xi)]/\sqrt{2}$ and $\nu = \mathbb{E}[f^2(\xi)]$. The parameters a_1, ν determine the limiting spectral measure $\mathcal{L}$ of $\mathbf{K}$ (since $\mathbf{K}$ and $\mathbf{K}_N$ asymptotically differ by a rank-4 matrix, they share the same limiting spectral measure) while a_2, a_2 determine the low rank structure within $\tilde{\mathbf{K}}_I$.

As an immediate consequence, arbitrary (square-summable) kernel functions f (with $a_0 = 0$) are asymptotically *equivalent* to the simple cubic function $\tilde{f}(x) = c_3 x^3 + c_2 x^2 + c_1 x - c_2$ having the *same* Hermite polynomial coefficients a_1, a_2, ν.[10]

The idea of this section is to design a prototypical family $\mathcal{F}$ of functions f having (i) universal properties with respect to (a_1, a_2, ν), i.e., for each (a_1, a_2, ν) there exists $f \in \mathcal{F}$ with these Hermite coefficients and (ii) having numerically advantageous properties. Thus, any arbitrary kernel function f can be mapped, through (a_1, a_2, ν), to a function in $\mathcal{F}$ with good numerical properties.

[10]The coefficients being related through $a_1 = 3c_3 + c_1$, $a_2 = \sqrt{2}c_2$ and $\nu = (3c_3 + c_1)^2 + 6c_3^2 + 2c_2^2$.

Table 1: Storage size and top eigenvector running time of $\mathbf{K}$ for piecewise constant and cubic f, in the setting of Figure 2 and 4.

f	Size (Mb)	Running time (s)
Piecewise	4.15	0.2390
Cubic	16.75	0.4244

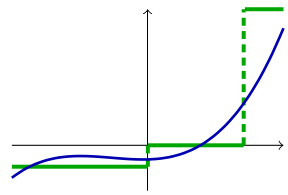


Figure 3: Piecewise constant (green) versus cubic (blue) function with equal (a_1, a_2, ν).

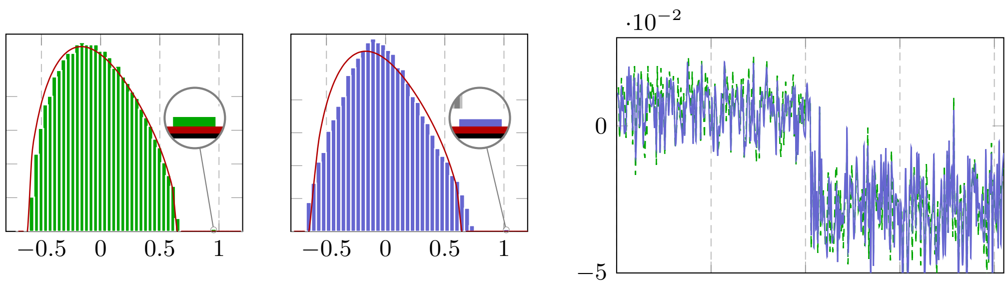
One such prototypical family $\mathcal{F}$ can be the set of f, parametrized by (t, s_-, s_+), and defined as

$$f(x) = \begin{cases} -rt & x \leq \sqrt{2}s_- \\ 0 & \sqrt{2}s_- < x \leq \sqrt{2}s_+ \\ t & x > \sqrt{2}s_+ \end{cases} \text{ with } \begin{cases} a_1 = \frac{t}{\sqrt{2\pi}}(e^{-s_+^2} + re^{-s_-^2}) \\ a_2 = \frac{t}{\sqrt{2\pi}}(s_+e^{-s_+^2} + rs_-e^{-s_-^2}) \\ \nu = \frac{t^2}{2}\left(1 - \operatorname{erf}(s_+)\right)(1+r) \end{cases} \tag{10}$$

where $r \equiv \frac{1-\operatorname{erf}(s_+)}{1+\operatorname{erf}(s_-)}$. Figure 3 displays f given in (10) together with the cubic function $c_3x^3 + c_2(x^2-1) + c_1x$ sharing the same Hermite coefficients (a_1, a_2, ν).

The class of equivalence of kernel functions induced by this mapping is quite unlike that raised in [EK10b] or [CBG16] in the "improper" scaling $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p)$ regime. While in the latter, functions $f(x)$ of the same class of equivalence are those having common $f'(0)$ and $f''(0)$ values, in the present case, these functions may have no similar local behavior (as shown in the example of Figure 3).

For the piecewise constant function defined in (10) and the associated cubic function having the same (a_1, a_2, ν), a close match is observed for both eigenvalues and top eigenvectors of $\mathbf{K}$ in Figure 4, with gains in both storage size and computational time displayed in Table 1.



(a) Piecewise constant

(b) Cubic polynomial

(c) Cubic (blue) versus piecewise (green) function

Figure 4: Eigenvalue distribution and top eigenvectors of $\mathbf{K}$ for the piecewise constant function (in green) and the associated cubic function (in blue) with the same (a_1, a_2, ν), performed on Bernoulli distribution with zero mean and unit variance, in the setting of Figure 2.

5 Concluding remarks

We have shown that inner-product kernel matrices $\sqrt{p}\mathbf{K} = f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})$ with $\mathbf{x}_i = \boldsymbol{\mu}_a + (\mathbf{I}_p + \mathbf{E}_a)^{\frac{1}{2}}\mathbf{z}_i$, $a \in \{1, 2\}$, asymptotically behave as a spiked random matrix model which spectrally only depends on three defining parameters of f. Turning $\mathbf{I}_p$ into a generic $\mathbf{C}$ covariance is more technically challenging, breaking most of the orthogonality properties of the orthogonal polynomial approach of the proofs, but a needed extension of the result.

Interestingly, this study can be compared to analyses in neural networks (see e.g., [PW17, BP19]) where it has been shown that in the case of sub-Gaussian entries for both random layer $\mathbf{W}$ and input data $\mathbf{X}$ the (limiting) spectrum of the Gram matrix $f(\mathbf{WX})f(\mathbf{WX})^\mathsf{T}$ (f understood entry-wise) is uniquely determined by the same (a_1, ν) coefficients. Our results may then be adapted to an improved understanding of classification in random neural networks.

References

[AAR00] George E Andrews, Richard Askey, and Ranjan Roy. *Special functions*, volume 71. Cambridge university press, 2000.

[ABDH+18] Hassan Ashtiani, Shai Ben-David, Nicholas Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Nearly tight sample complexity bounds for learning mixtures of gaussians via sample compression schemes. In *Advances in Neural Information Processing Systems*, pages 3412–3421, 2018.

[AS65] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1965.

[BAP05] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[BGN11] Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

[Bil12] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, third edition, 2012.

[BP19] Lucas Benigni and Sandrine Péché. Eigenvalue distribution of nonlinear models of random matrices. *arXiv preprint arXiv:1904.03090*, 2019.

[CBG16] Romain Couillet and Florent Benaych-Georges. Kernel spectral clustering of large dimensional data. *Electronic Journal of Statistics*, 10(1):1393–1454, 2016.

[CLM18] Romain Couillet, Zhenyu Liao, and Xiaoyi Mai. Classification asymptotics in the random matrix regime. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1875–1879. IEEE, 2018.

[CS13] Xiuyuan Cheng and Amit Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 2(04):1350010, 2013.

[DV13] Yen Do and Van Vu. The spectrum of random kernel matrices: universality results for rough and varying kernels. *Random Matrices: Theory and Applications*, 2(03):1350005, 2013.

[EK10a] Noureddine El Karoui. On information plus noise kernel random matrices. *The Annals of Statistics*, 38(5):3191–3216, 2010.

[EK10b] Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, 2010.

[FM19] Zhou Fan and Andrea Montanari. The spectral norm of random inner-product kernel matrices. *Probability Theory and Related Fields*, 173(1-2):27–85, 2019.

[HJ12] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, second edition, 2012.

[JL09] Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

[KMV16] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Disentangling gaussians. *Communications of the ACM*, 55, 2016.

[LC18] Zhenyu Liao and Romain Couillet. On the spectrum of random features maps of high dimensional data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3063–3071. PMLR, 2018.

[LC19] Zhenyu Liao and Romain Couillet. A large dimensional analysis of least squares support vector machines. *IEEE Transactions on Signal Processing*, 67(4):1065–1074, 2019.

[LLC18] Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

[NJW02] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

[PW17] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, pages 2637–2646, 2017.

[Rud64] Walter Rudin. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

[Wil97] Christopher KI Williams. Computing with infinite networks. In *Advances in neural information processing systems*, pages 295–301, 1997.

Supplementary Material

Inner-product Kernels are Asymptotically Equivalent to Binary Discrete Kernels

A The non-trivial classification regime

In the ideal case where $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\mathbf{E}_1, \mathbf{E}_2$ are perfectly known, the (decision optimal) Neyman-Pearson test to decide on the class of an unknown and normally distributed $\mathbf{x}$, genuinely belonging to $\mathcal{C}_1$, consists in the following comparison

$$(\mathbf{x}-\boldsymbol{\mu}_2)^\mathsf{T}(\mathbf{I}_p+\mathbf{E}_2)^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)-(\mathbf{x}-\boldsymbol{\mu}_1)^\mathsf{T}(\mathbf{I}_p+\mathbf{E}_1)^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\gtrless}} \log\frac{\det(\mathbf{I}_p+\mathbf{E}_1)}{\det(\mathbf{I}_p+\mathbf{E}_2)}.$$

Writing $\mathbf{x}=\boldsymbol{\mu}_1+(\mathbf{I}_p+\mathbf{E}_1)^{\frac{1}{2}}\mathbf{z}$ so that $\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_p)$, the above test is then equivalent to

$$\begin{aligned}T(\mathbf{x})&=\frac{1}{p}\mathbf{z}^\mathsf{T}\left((\mathbf{I}_p+\mathbf{E}_1)^{\frac{1}{2}}(\mathbf{I}_p+\mathbf{E}_2)^{-1}(\mathbf{I}_p+\mathbf{E}_1)^{\frac{1}{2}}-\mathbf{I}_p\right)\mathbf{z}+\frac{2}{p}\Delta\boldsymbol{\mu}^\mathsf{T}(\mathbf{I}_p+\mathbf{E}_2)^{-1}(\mathbf{I}_p+\mathbf{E}_1)^{\frac{1}{2}}\mathbf{z}\\&\quad+\frac{1}{p}\Delta\boldsymbol{\mu}^\mathsf{T}(\mathbf{I}_p+\mathbf{E}_2)^{-1}\Delta\boldsymbol{\mu}-\frac{1}{p}\log\frac{\det(\mathbf{I}_p+\mathbf{E}_1)}{\det(\mathbf{I}_p+\mathbf{E}_2)}\underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\gtrless}}0\end{aligned}$$

where $\Delta\boldsymbol{\mu}\equiv\boldsymbol{\mu}_1-\boldsymbol{\mu}_2$. Since, for $\mathbf{U}\in\mathbb{R}^{p\times p}$ an eigenvector basis for $(\mathbf{I}_p+\mathbf{E}_1)^{\frac{1}{2}}(\mathbf{I}_p+\mathbf{E}_2)^{-1}(\mathbf{I}_p+\mathbf{E}_1)^{\frac{1}{2}}-\mathbf{I}_p$, $\mathbf{U}\mathbf{z}$ has the same distribution as $\mathbf{z}$, with a careful application of the Lyapunov's central limit theorem (see for example [Bil12, Theorem 27.3]), along with the assumption $\|\mathbf{E}_a\|=o(1)$ for $a\in\{1,2\}$, we obtain

$$V_T^{-\frac{1}{2}}(T(\mathbf{x})-E_T)\xrightarrow{\mathcal{L}}\mathcal{N}(0,1)$$

as $p\to\infty$, with

$$\begin{aligned}E_T&\equiv\frac{1}{p}\operatorname{tr}\left((\mathbf{I}_p+\mathbf{E}_1)(\mathbf{I}_p+\mathbf{E}_2)^{-1}\right)-1+\frac{1}{p}\Delta\boldsymbol{\mu}^\mathsf{T}(\mathbf{I}_p+\mathbf{E}_2)^{-1}\Delta\boldsymbol{\mu}-\frac{1}{p}\log\frac{\det(\mathbf{I}_p+\mathbf{E}_1)}{\det(\mathbf{I}_p+\mathbf{E}_2)}\\V_T&\equiv\frac{2}{p^2}\|(\mathbf{I}_p+\mathbf{E}_1)^{\frac{1}{2}}(\mathbf{I}_p+\mathbf{E}_2)^{-1}(\mathbf{I}_p+\mathbf{E}_1)^{\frac{1}{2}}-\mathbf{I}_p\|_F^2\\&\quad+\frac{4}{p^2}\Delta\boldsymbol{\mu}^\mathsf{T}(\mathbf{I}_p+\mathbf{E}_2)^{-1}(\mathbf{I}_p+\mathbf{E}_1)(\mathbf{I}_p+\mathbf{E}_2)^{-1}\Delta\boldsymbol{\mu}.\end{aligned}$$

The classification error rate is thus non-trivial (i.e., converging neither to 0 not 1 as $p\to\infty$) if E_T and $\sqrt{V_T}$ are of the same order of magnitude (with respect to p).

In the case where $\mathbf{E}_1=\mathbf{E}_2=\mathbf{E}$,

$$E_T=\frac{1}{p}\Delta\boldsymbol{\mu}^\mathsf{T}(\mathbf{I}_p+\mathbf{E})^{-1}\Delta\boldsymbol{\mu}=O(\|\Delta\boldsymbol{\mu}\|^2p^{-1}),\ \sqrt{V_T}=\frac{2}{p}\sqrt{\Delta\boldsymbol{\mu}^\mathsf{T}(\mathbf{I}_p+\mathbf{E})^{-1}\Delta\boldsymbol{\mu}}=O(\|\Delta\boldsymbol{\mu}\|p^{-1})$$

so that we must as least demand $\|\Delta\boldsymbol{\mu}\|\geq O(1)$ (which, up to centering, is equivalent to asking $\|\boldsymbol{\mu}_a\|\geq O(1)$ for $a\in\{1,2\}$).

Under the critical condition $\|\Delta\boldsymbol{\mu}\|=O(1)$, we move on to the study of the condition on the covariance $\mathbf{E}_a$. To this end, a Taylor expansion can be performed for $\mathbf{I}_p+\mathbf{E}_2$ around $\mathbf{I}_p+\mathbf{E}_1$ so that

$$\begin{aligned}E_T&=\frac{1}{p}\Delta\boldsymbol{\mu}^\mathsf{T}(\mathbf{I}_p+\mathbf{E}_1)^{-1}\Delta\boldsymbol{\mu}+\frac{1}{2p}\|(\mathbf{I}_p+\mathbf{E}_1)^{-1}\Delta\mathbf{E}\|_F^2+o(p^{-1})\\V_T&=\frac{2}{p^2}\|(\mathbf{I}_p+\mathbf{E}_1)^{-1}\Delta\mathbf{E}\|_F^2+\frac{4}{p^2}\Delta\boldsymbol{\mu}^\mathsf{T}(\mathbf{I}_p+\mathbf{E}_1)^{-1}\Delta\boldsymbol{\mu}+o(p^{-2}).\end{aligned}$$

with $\Delta\mathbf{E}\equiv\mathbf{E}_1-\mathbf{E}_2$. Thus one must have $\|\Delta\mathbf{E}\|\geq O(p^{-1/2})$ for $\|(\mathbf{I}_p+\mathbf{E}_1)^{-1}\Delta\mathbf{E}\|_F^2$ not to vanish for p large and for $\Delta\mathbf{E}$ to have discriminating power. It is thus convenient to request $\|\mathbf{E}_a\|\geq O(p^{-1/2})$ for $a\in\{1,2\}$, which unfolds from

$$|\operatorname{tr}\mathbf{E}_a|\geq O(\sqrt{p}),\quad\text{or}\quad\|\mathbf{E}_a\|_F^2\geq O(1).$$

Yet, as noticed in [CLM18], many classification algorithms, either supervised, semi-supervised or unsupervised, are not able to achieve the optimal rate $\|\mathbf{E}_a\|_F^2=O(1)$ when n,p are of the same order of magnitude. Indeed, the best possible rate $\|\mathbf{E}_a\|_F^2=O(\sqrt{p})$ can only be obtained in very particular cases, for instance if $|\operatorname{tr}\mathbf{E}_a|=o(\sqrt{p})$ and $\|\boldsymbol{\mu}_a\|=o(1)$ as investigated in [LC19]. This thus leads to the non-trivial classification condition demanded in Assumption 1.

B Exact computation of ϕ in the Gaussian case

In the Gaussian case where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ we resort to computing, as in [Wil97, LLC18], the integral

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}}[(\mathbf{z}^\mathsf{T}\mathbf{a})^{k_1}(\mathbf{z}^\mathsf{T}\mathbf{b})^{k_2}] = (2\pi)^{-p/2}\int_{\mathbb{R}^p}(\mathbf{z}^\mathsf{T}\mathbf{a})^{k_1}(\mathbf{z}^\mathsf{T}\mathbf{b})^{k_2}e^{-\|\mathbf{z}\|^2/2}d\mathbf{z}\\
&= \frac{1}{2\pi}\int_{\mathbb{R}^2}(\tilde z_1\tilde a_1)^{k_1}(\tilde z_1\tilde b_1+\tilde z_2\tilde b_2)^{k_2}e^{-(\tilde z_1^2+\tilde z_2^2)/2}d\tilde z_1 d\tilde z_2 = \frac{1}{2\pi}\int_{\mathbb{R}^2}(\tilde{\mathbf{z}}^\mathsf{T}\tilde{\mathbf{a}})^{k_1}(\tilde{\mathbf{z}}^\mathsf{T}\tilde{\mathbf{b}})^{k_2}e^{-\|\tilde{\mathbf{z}}\|^2/2}d\tilde{\mathbf{z}}
\end{aligned}$$

where we apply the Gram-Schmidt procedure to project $\mathbf{z}$ into the two-dimensional space[11] spanned by $\mathbf{a}, \mathbf{b}$ with $\tilde a_1 = \|\mathbf{a}\|$, $\tilde b_1 = \frac{\mathbf{a}^\mathsf{T}\mathbf{b}}{\|\mathbf{a}\|}$, $\tilde b_2 = \sqrt{\|\mathbf{b}\|^2 - \frac{(\mathbf{a}^\mathsf{T}\mathbf{b})^2}{\|\mathbf{a}\|^2}}$ and denote $\tilde{\mathbf{z}} = [\tilde z_1; \tilde z_2]$, $\tilde{\mathbf{a}} = [\tilde a_1; 0]$ and $\tilde{\mathbf{b}} = [\tilde b_1; \tilde b_2]$. As a consequence, we obtain, for k even,

$$\begin{aligned}
&\mathbb{E}\left[(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})^k\right] = \mathbb{E}[\xi^k] = (k-1)!!;\\
&\mathbb{E}_{\mathbf{z}_i}\left[(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})^k(\mathbf{z}_i^\mathsf{T}\mathbf{b})\right] = \mathbb{E}_{\mathbf{z}_i}\left[(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})^{k-1}(\mathbf{z}_i^\mathsf{T}\mathbf{b})^2\right] = 0;\\
&\mathbb{E}_{\mathbf{z}_i}\left[(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})^{k-1}(\mathbf{z}_i^\mathsf{T}\mathbf{b})\right] = (k-1)!!(\|\mathbf{z}_j\|/\sqrt{p})^{k-2}(\mathbf{z}_j^\mathsf{T}\mathbf{b})/\sqrt{p};\\
&\mathbb{E}_{\mathbf{z}_i}\left[(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})^k(\mathbf{z}_i^\mathsf{T}\mathbf{b})^2\right] = (k-1)!!\left(k(\|\mathbf{z}_j\|/\sqrt{p})^{k-2}(\mathbf{z}_j^\mathsf{T}\mathbf{b}/\sqrt{p})^2 + (\|\mathbf{z}_j\|/\sqrt{p})^k\|\mathbf{b}\|^2\right);
\end{aligned}$$

where we denote $k!!$ the double factorial of an integral k such that $k!! = k(k-2)(k-4)\cdots$. This futher leads to, in the Gaussian case, the expression of $\tilde{\mathbf{K}}_I$ in Proposition 2.

C Proof of Proposition 2

Define by $\mathbf{L}$ the matrix with $\mathbf{L}_{ij} \equiv [\frac{1}{p}(\mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T} + \mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{Z} + \mathbf{Z}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T})]_{ij}$ for $i \neq j$ and $\mathbf{L}_{ii} = 0$. Then $\mathbf{K}_I$ can be written as

$$\begin{aligned}
\mathbf{K}_I &= k(\mathbf{Z}^\mathsf{T}\mathbf{Z}/\sqrt{p})^{\circ(k-1)} \circ \mathbf{L} + \mathbf{\Phi},\\
\mathbf{\Phi}_{ij} &\equiv \frac{k}{p}(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})^{k-1}\mathbf{z}_i^\mathsf{T}\left(\frac{1}{2}(\mathbf{E}_a+\mathbf{E}_b) - \frac{1}{8}(\mathbf{E}_a-\mathbf{E}_b)^2\right)\mathbf{z}_j\\
&\quad + \frac{k(k-1)}{8p}(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})^{k-2}\frac{1}{\sqrt{p}}(\mathbf{z}_i^\mathsf{T}(\mathbf{E}_a+\mathbf{E}_b)\mathbf{z}_j)^2
\end{aligned}$$

for $i \neq j$ and $\mathbf{\Phi}_{ii} = 0$. With this expression, the proof of Proposition 2 can be divided into three steps.

Concentration of $\mathbf{\Phi}$. We first show that, $\|\mathbf{\Phi} - \mathbb{E}[\mathbf{\Phi}]\| \to 0$ almost surely, as $n, p \to \infty$. This follows from the observation that $\mathbf{\Phi}$ is a $p^{-1/4}$ rescaling (since $\|\mathbf{E}_a\| = O(p^{-1/4})$) of the null model $\mathbf{K}_N$, which concentrates around its expectation in the sense that $\|\mathbf{K}_N - \mathbb{E}[\mathbf{K}_N]\| = O(1)$ for $\mathbb{E}[\mathbf{K}_N] = O(\sqrt{p})$ if $a_0 \neq 0$ (see Remark 1). Indeed, it is shown in [FM19, Theorem 1.7] that, the leading eigenvalue of order $O(\sqrt{p})$ discarded (arising from $\mathbb{E}[\mathbf{K}_N]$), $\mathbf{K}_N$ is of bounded operator norm for all large n, p with probability one; this, together with the fact that $\|\mathbb{E}[\mathbf{\Phi}]\| = O(1)$ that will be shown subsequently, allows us to conclude that $\|\mathbf{\Phi} - \mathbb{E}[\mathbf{\Phi}]\| \to 0$ as $n, p \to \infty$.

Computation of $\mathbb{E}[\mathbf{\Phi}]$: beyond the Gaussian case. We then show that, for random vectors $\mathbf{z}$ with zero mean, unit variance and bounded moments entries, the expression of $\mathbb{E}[\mathbf{\Phi}]$ coincides with the Gaussian case. To this end, recall that the entries of $\mathbf{\Phi}$ are the sum of random variables of the type

$$\phi = \frac{C}{\sqrt{p}}(\mathbf{x}^\mathsf{T}\mathbf{y}/\sqrt{p})^\alpha(\mathbf{x}^\mathsf{T}\mathbf{F}\mathbf{y})^\beta$$

for independent random vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ with i.i.d. zero mean, unit variance and finite moments (uniformly on p) entries, deterministic $\mathbf{F} \in \mathbb{R}^{p\times p}$, $C \in \mathbb{R}$, $\alpha \in \mathbb{N}$ and $\beta \in \{1, 2\}$. Let us start with the case $\beta = 1$ and expand ϕ as

$$\phi = \frac{C}{\sqrt{p}}\left(\frac{1}{\sqrt{p}}\sum_{i_1=1}^p x_{i_1}y_{i_1}\right)\cdots\left(\frac{1}{\sqrt{p}}\sum_{i_\alpha=1}^p x_{i_\alpha}y_{i_\alpha}\right)\left(\sum_{j_1,j_2=1}^p F_{j_1,j_2}x_{j_1}y_{j_2}\right) \tag{11}$$

[11] By assuming first that $\mathbf{a}, \mathbf{b}$ are linearly independent before extending by continuity to $\mathbf{a}, \mathbf{b}$ proportional.

with x_i and y_i the i-th entry of $\mathbf{x}$ and $\mathbf{y}$, respectively, so that (i) x_i is independent of y_j for all i, j and (ii) x_i is independent of x_j for $i \neq j$ with $\mathbb{E}[x_i] = 0$, $\mathbb{E}[x_i^2] = 1$ and $\mathbb{E}[|x_i|^k] \leq C_k$ for some C_k independent of p (and similarly for $\mathbf{y}$).

At this point, note that to ensure $\mathbb{E}[\mathbf{K}_I]$ has non-vanishing operator norm as $n, p \to \infty$, we need $\mathbb{E}[\phi] \geq O(p^{-1})$ since $\|\mathbf{A}\| \leq p\|\mathbf{A}\|_\infty$ for $\mathbf{A} \in \mathbb{R}^{p \times p}$. Also, note that (as $\beta = 1$), all terms in the sum $\sum_{j_1,j_2=1}^{p} F_{j_1,j_2} x_{j_1} y_{j_2}$ with $j_1 \neq j_2$ must be zero since in other terms x_i always appears together with y_i, so that all terms with $j_1 \neq j_2$ give rise to zero in expectation. Hence, the p^2 terms of the sum only contain p nonzero terms in expectation (those with $j_1 = j_2$). The arbitrary (absolute) moments of x and y being finite, the first αp terms must be divided into $\lceil \alpha \rceil / 2$ groups of size two (containing $O(p)$ terms) so that, with the normalization by p^{-1} for each group of size two, the associated expectation is not vanishing. We shall thus discuss the following two cases:

1. α even: the α terms in the sum form $\alpha/2$ groups with different indices each and also different from $j_1 = j_2$. Therefore we have $\mathbb{E}_{x_j}[\phi] = 0$ and $\mathbb{E}[\phi] = 0$.
2. α odd: the α terms in the sum form $(\alpha - 1)/2$ groups with indices different from each other and the remaining one goes with the last term containing $\mathbf{F}$ and one has $\mathbb{E}[\phi] = \frac{C\alpha!!}{p} \operatorname{tr}(\mathbf{F})$ by a simple combinatorial argument.

The case $\beta = 2$ follows exactly the same line of arguments except that j_1 may not equal j_2 to give rise to non-vanishing terms.

Concentration of Hadamard product. It now remains to treat the term $k(\mathbf{Z}^\mathsf{T}\mathbf{Z}/\sqrt{p})^{\circ(k-1)} \circ \mathbf{L}$ and show it also has an asymptotically deterministic behavior (as $\mathbf{\Phi}$). It can be shown that

$$\|\mathbf{N} \circ \mathbf{L}\| \to 0, \quad n, p \to \infty$$

with $\mathbf{N} = (\mathbf{Z}^\mathsf{T}\mathbf{Z}/\sqrt{p})^{\odot(k-1)} - (k-2)!!\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$ for k odd and $\mathbf{N} = (\mathbf{Z}^\mathsf{T}\mathbf{Z}/\sqrt{p})^{\odot(k-1)}$ for k even.

To prove this, note that, depending on the key parameter $a_0 = \mathbb{E}[f(\xi)]$, the operator norm of $f(\mathbf{Z}^\mathsf{T}\mathbf{Z}/\sqrt{p})/\sqrt{p}$ is either of order $O(\sqrt{p})$ for $a_0 \neq 0$ or $O(1)$ for $a_0 = 0$. In particular, for monomial $f(x) = x^k$ under study here, we have $a_0 = \mathbb{E}[\xi^k] = 0$ for k odd and $a_0 = \mathbb{E}[\xi^k] = (k-1)!! \neq 0$ for k even, $\xi \sim \mathcal{N}(0, 1)$. To control the operator norm of the Hadamard product between matrices, we introduce the following lemma.

Lemma 1. *For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p \times p}$, we have $\|\mathbf{A} \circ \mathbf{B}\| \leq \sqrt{p}\|\mathbf{A}\|_\infty \|\mathbf{B}\|$.*

Proof of Lemma 1. Let $\mathbf{e}_1, \ldots, \mathbf{e}_p$ be the canonical basis vectors of $\mathbb{R}^p$, then for all $1 \leq i \leq p$,

$$\|(\mathbf{A} \circ \mathbf{B})\mathbf{e}_i\| \leq \max_{i,j} |\mathbf{A}_{ij}| \|\mathbf{B}\mathbf{e}_i\| = \|\mathbf{A}\|_\infty \|\mathbf{B}\mathbf{e}_i\| \leq \|\mathbf{A}\|_\infty \|\mathbf{B}\|.$$

As a consequence, for any $\mathbf{v} = \sum_{i=1}^p v_i \mathbf{e}_i$, we obtain

$$\|(\mathbf{A} \circ \mathbf{B})\mathbf{v}\| \leq \sum_{i=1}^p |v_i| \|(\mathbf{A} \circ \mathbf{B})\mathbf{e}_i\| \leq \sum_{i=1}^p |v_i| \|\mathbf{A}\|_\infty \|\mathbf{B}\|$$

which, by Cauchy-Schwarz inequality further yields $\sum_{i=1}^p |v_i| \leq \sqrt{p}\|\mathbf{v}\|$. This concludes the proof of Lemma 1. □

Lemma 1 tells us that the Hadamard product between a matrix with $o(p^{-1/2})$ entry and a matrix with bounded operator norm is of vanishing operator norm, as $p \to \infty$. As such, since $\|\mathbf{N}\| = O(1)$ and $\mathbf{L}$ has $O(p^{-1})$ entries, we have $\|\mathbf{N} \circ \mathbf{L}\| \to 0$. This concludes the proof of Proposition 2.

D Proof of Theorem 2

The proof follows from the fact that the individual coefficients of the Hermite polynomials $P_\kappa(x) = \sum_{l=0}^{\kappa} c_{\kappa,l} x^l$ satisfy the following recurrent relation [AS65]

$$c_{\kappa+1,l} = \begin{cases} -\kappa c_{\kappa-1,l} & l = 0; \\ c_{\kappa,l-1} - \kappa c_{\kappa-1,l} & l \geq 1; \end{cases} \tag{12}$$

with $c_{0,0} = 1$, $c_{1,0} = 0$ and $c_{1,1} = 1$. As a consequence, by indexing the informative matrix in Proposition 2 of the monomial $f(x) = x^l$ as $\tilde{\mathbf{K}}_{I,l}$, we have for odd $\kappa \geq 3$,

$$\tilde{\mathbf{K}}_I = \sum_{l=1,3,\ldots}^{\kappa} c_{\kappa,l}\tilde{\mathbf{K}}_{I,l} = \sum_{l=1,3,\ldots}^{\kappa} c_{\kappa,l} l!!(\mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T} + \mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{Z} + \mathbf{Z}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T})/p - \mathrm{diag}(\cdot) = \mathbf{0}$$

with $[\mathbf{X} - \mathrm{diag}(\cdot)]_{ij} = \mathbf{X}_{ij}\delta_{i\neq j}$. This follows from the fact that, for $\kappa \geq 3$, we have both $\sum_{l=1,3,\ldots}^{\kappa} c_{\kappa,l} l!! = 0$ and $\sum_{l=0,2,\ldots}^{\kappa+1} c_{\kappa+1,l}(l+1)!! = 0$. The latter is proved by induction on κ: first, for $\kappa = 3$, we have $c_{3,1} + 3c_{3,3} = c_{4,0} + 3c_{4,2} + 15c_{4,4} = 0$; then, assuming κ odd, we have $\sum_{l=1,3,\ldots}^{\kappa} c_{\kappa,l} l!! = \sum_{l=0,2,\ldots}^{\kappa+1} c_{\kappa+1,l}(l+1)!! = 0$ so that, together with (12)

$$\sum_{l=1,3,\ldots}^{\kappa+2} c_{\kappa+2,l} l!! = \sum_{l=1,\ldots}^{\kappa+2} (c_{\kappa+1,l-1} - (\kappa+1)c_{\kappa,l}) l!! = \sum_{l=1,\ldots}^{\kappa+2} c_{\kappa+1,l-1} l!! = \sum_{l=0,2,\ldots}^{\kappa+1} c_{\kappa+1,l}(l+1)!! = 0$$

as well as

$$\sum_{l=0,2,\ldots}^{\kappa+3} c_{\kappa+3,l}(l+1)!! = -(\kappa+2)c_{\kappa+1,0} + \sum_{l=2,4,\ldots}^{\kappa+3} (c_{\kappa+2,l-1} - (\kappa+2)c_{\kappa+1,l})(l+1)!! = 0$$

where we used $c_{\kappa,l} = 0$ for $l \geq \kappa+1$. Similar arguments hold for the case of κ even, which concludes the proof.