# On the Spectrum of Random Features Maps of High Dimensional Data

**Anonymous Authors**[1]

## Abstract

Random feature maps are ubiquitous in modern statistical machine learning, where they generalize random projections by means of powerful, yet often difficult to analyze nonlinear operators. In this paper, we leverage the "concentration" phenomenon induced by random matrix theory to perform a spectral analysis on the Gram matrix of these random feature maps, here for Gaussian mixture models of simultaneously large dimension and size. Our results are instrumental to a deeper understanding on the interplay of the nonlinearity and the statistics of the data, thereby allowing for a better tuning of random feature-based techniques.

## 1. Introduction

Finding relevant features is one of the key steps for solving a machine learning problem. To this end, the backpropagation algorithm is probably the best-known method, with which superhuman performances are commonly achieved for specific tasks in applications such as computer vision (Krizhevsky et al., 2012) and many others (Schmidhuber, 2015). But data-driven approaches such as the backpropagation method, in addition to being computationally demanding, fail to cope with limited amounts of available training data.

One successful alternative in this regard is the use of "random features", exploited both in feed-forward neural networks (Huang et al., 2012; Scardapane & Wang, 2017), in large-scale kernel estimation (Rahimi & Recht, 2008; Vedaldi & Zisserman, 2012) and more recently in random sketching schemes (Keriven et al., 2016). Random feature maps consist in projections randomly exploring the set of nonlinear representations of the data, hopefully extracting features relevant to some given task. The nonlinearities make these representations more mighty but meanwhile theoretically more difficult to analyze and optimize.

Infinitely large random features maps are nonetheless well understood as they result in (asymptotically) equivalent kernels, the most popular example being random Fourier features and their limiting radial basis kernels (Rahimi & Recht, 2008). Beyond those asymptotic results, recent advances in random matrix theory give rise to unexpected simplification on the understanding of the finite-dimensional version of these kernels, i.e., when the data number and size are large but of similar order as the random feature vector size (El Karoui et al., 2010; Couillet et al., 2016). Following the same approach, in this work, we perform a spectral analysis on the Gram matrix of the random feature matrices. This matrix is of key relevance in many associated machine learning methods (e.g., spectral clustering (Ng et al., 2002) and kernel SVM (Schölkopf & Smola, 2002)) and understanding its spectrum casts an indispensable light on their asymptotic performances. In the remainder of the article, we shall constantly consider spectral clustering as a concrete example of application; however, similar analyses can be performed for other types of random feature-based algorithms.

Our contribution is twofold. From a random matrix theory perspective, it is a natural extension of the sample covariance matrix analysis (Silverstein & Bai, 1995) to a nonlinear setting and can also be seen as the generalization of the recent work of (Pennington & Worah, 2017) to a more practical data model. From a machine learning point of view, we describe quantitatively the mutual influence of different nonlinearities and data statistics on the resulting random feature maps. More concretely, based on the ratio of two coefficients from our analysis, commonly used activation functions are divided into three classes: means-oriented, covariance-oriented and balanced, which eventually allows one to choose the activation function with respect to the statistical properties of the data (or task) at hand, with a solid theoretical basis.

We show by experiments that our results, applicable theoretically only to Gaussian mixture data, show an almost perfect match when applied to some real-world datasets. We are thus optimistic that our findings, although restricted to Gaussian assumptions on the data model, can be applied to a larger set of problems beyond strongly structured ones.

*Notations*: Boldface lowercase (uppercase) characters stand

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

for vectors (matrices), and non-boldface scalars respectively. $\mathbf{1}_T$ is the column vector of ones of size $T$, and $\mathbf{I}_T$ the $T \times T$ identity matrix. The notation $(\cdot)^\mathsf{T}$ denotes the transpose operator. The norm $\|\cdot\|$ is the Euclidean norm for vectors and the operator norm for matrices.

In the remainder of this article, we introduce the objects of interest and necessary preliminaries in Section 2. Our main results on the spectrum of random feature maps will be presented and discussed in Section 3, followed by experiments on two types of classification tasks in Section 4. The article closes on concluding remarks and envisioned extensions in Section 5.

## 2. Problem Statement and Preliminaries

Let $\mathbf{x}_1, \ldots, \mathbf{x}_T \in \mathbb{R}^p$ be independent data vectors, each belonging to one of $K$ distribution classes $\mathcal{C}_1, \ldots, \mathcal{C}_K$. Class $\mathcal{C}_a$ has cardinality $T_a$, for all $a \in \{1, \ldots, K\}$. We assume that the data vector $\mathbf{x}_i$ follows a Gaussian mixture model[1], i.e.,

$$\mathbf{x}_i = \frac{1}{\sqrt{p}} \boldsymbol{\mu}_a + \boldsymbol{\omega}_i$$

with $\boldsymbol{\omega}_i \sim \mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{C}_a)$ for some mean $\boldsymbol{\mu}_a \in \mathbb{R}^p$ and covariance $\mathbf{C}_a \in \mathbb{R}^{p \times p}$ of associated class $\mathcal{C}_a$.

We denote the data matrix $\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_T\right] \in \mathbb{R}^{p \times T}$ of size $T$ by cascading all $\mathbf{x}_i$ as column vectors. To extract random features, $\mathbf{X}$ is premultiplied by some random matrix $\mathbf{W} \in \mathbb{R}^{n \times p}$ with i.i.d. entries and then applied entry-wise some nonlinear *activation function* $\sigma(\cdot)$ to obtain the random feature matrix $\boldsymbol{\Sigma} \equiv \sigma(\mathbf{W}\mathbf{X}) \in \mathbb{R}^{n \times T}$, whose columns are simply $\sigma(\mathbf{W}\mathbf{x}_i)$ the associated random feature of $\mathbf{x}_i$.

In this article, we focus on the Gram matrix $\mathbf{G} \equiv \frac{1}{n}\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma}$ of the random features, the entry $(i, j)$ of which is given by

$$\mathbf{G}_{ij} = \frac{1}{n}\sigma(\mathbf{W}\mathbf{x}_i)^\mathsf{T}\sigma(\mathbf{W}\mathbf{x}_j) = \frac{1}{n}\sum_{k=1}^{n} \sigma(\mathbf{w}_k^\mathsf{T}\mathbf{x}_i)\sigma(\mathbf{w}_k^\mathsf{T}\mathbf{x}_j)$$

with $\mathbf{w}_k^\mathsf{T}$ the $k$-th row of $\mathbf{W}$. Note that all $\mathbf{w}_k$ follow the same distribution, so that taking expectation over $\mathbf{w} \equiv \mathbf{w}_k$ of the above equation one results in the average kernel matrix $\boldsymbol{\Phi}$, with

$$\boldsymbol{\Phi}_{ij} \equiv \boldsymbol{\Phi}(\mathbf{x}_i, \mathbf{x}_j) = \mathbb{E}_\mathbf{w}[\mathbf{G}_{ij}] = \mathbb{E}_\mathbf{w}[\sigma(\mathbf{w}^\mathsf{T}\mathbf{x}_i)\sigma(\mathbf{w}^\mathsf{T}\mathbf{x}_j)]. \tag{1}$$

When the entries of $\mathbf{W}$ follow a standard Gaussian distribution, one can compute the generic form $\boldsymbol{\Phi}(\mathbf{a}, \mathbf{b}) = \mathbb{E}_\mathbf{w}[\sigma(\mathbf{w}^\mathsf{T}\mathbf{a})\sigma(\mathbf{w}^\mathsf{T}\mathbf{b})]$ by applying the integral trick from (Williams, 1997), for a large set of nonlinear functions $\sigma(\cdot)$ and arbitrary vector $\mathbf{a}, \mathbf{b}$ of appropriate dimension. We list the results for commonly used functions in Table 1.

[1] We normalize the data by $\frac{1}{\sqrt{p}}$ to guarantee that $\|\mathbf{x}_i\| = O(1)$ with high probability when $\|\mathbf{C}_a\| = O(1)$.

Since the Gram matrix $\mathbf{G}$ describes the correlation of data in the *feature space*, it is natural to recenter $\mathbf{G}$, and thus $\boldsymbol{\Phi}$ by pre- and post-multiplying a projection matrix $\mathbf{P} \equiv \mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^\mathsf{T}$. In the case of $\boldsymbol{\Phi}$, we get

$$\boldsymbol{\Phi}_c \equiv \mathbf{P}\boldsymbol{\Phi}\mathbf{P}.$$

In the recent line of works (Louart et al., 2017; Pennington & Worah, 2017), it has been shown that the large dimensional (large $n, p, T$) characterization of $\mathbf{G}$, in particular its eigenspectrum, is fully determined by $\boldsymbol{\Phi}$ and the ratio $n/p$. For instance, by defining the *empirical spectral distribution* of $\mathbf{G}_c = \mathbf{P}\mathbf{G}\mathbf{P}$ as $\rho_{\mathbf{G}_c}(x) \equiv \frac{1}{T}\sum_{i=1}^{T} 1_{\lambda_i \leq x}(x)$, with $\lambda_1, \ldots, \lambda_T$ the eigenvalues of $\mathbf{G}_c$, it has been shown in (Louart et al., 2017) that, as $n, p, T \to \infty$, $\rho_\mathbf{G}(x)$ almost surely converges to a non-random distribution $\rho(x)$, referred to as the *limiting spectral distribution* of $\mathbf{G}_c$ such that

$$\rho(x) = \frac{1}{\pi}\lim_{y \to 0^+} \int_{-\infty}^{x} \Im\left[m(t + iy)\right] dt.$$

with $m(z)$ the associated Stieltjes transform, explicitly given by

$$m(z) = \frac{1}{n}\operatorname{tr}\left(\frac{\boldsymbol{\Phi}_c}{1 + \delta(z)} - z\mathbf{I}_T\right)^{-1}$$

with $\delta(z)$ the unique solution of

$$\delta(z) = \frac{1}{n}\operatorname{tr}\left(\boldsymbol{\Phi}_c\left(\frac{\boldsymbol{\Phi}_c}{1 + \delta(z)} - z\mathbf{I}_T\right)^{-1}\right).$$

As a consequence, in the objective of understanding the asymptotic behavior of $\mathbf{G}_c$ as $n, p, T$ are simultaneously large, we shall focus our analysis on $\boldsymbol{\Phi}_c$. To this end, the following assumptions will be needed throughout the paper.

**Assumption 1** (Growth rate)**.** *As $T \to \infty$,*

1. $\frac{p}{T} \to c_0 \in (0, \infty)$
2. *for each* $a \in \{1, \ldots, K\}$, $\frac{T_a}{T} \to c_a \in (0, 1)$
3. $\|\boldsymbol{\mu}_a\| = O(1)$
4. *letting* $\mathbf{C}^\circ \equiv \sum_{a=1}^{K} \frac{T_a}{T}\mathbf{C}_a$ *and for each* $a \in \{1, \ldots, K\}$, $\mathbf{C}_a^\circ \equiv \mathbf{C}_a - \mathbf{C}^\circ$, *then* $\|\mathbf{C}_a\| = O(1)$ *and* $\frac{1}{\sqrt{p}}\operatorname{tr}(\mathbf{C}_a^\circ) = O(1)$
5. *for technical convenience we assume in addition that* $\tau \equiv \frac{1}{p}\operatorname{tr}\left(\mathbf{C}^\circ\right)$ *converges in* $(0, \infty)$.

Assumption 1 ensures that the information about data means or covariances is neither too simple nor impossible to be extracted from the data, as closely investigated in (Couillet et al., 2016).

*Table 1.* $\mathbf{\Phi}(\mathbf{a},\mathbf{b})$ for different $\sigma(\cdot)$, $\angle(\mathbf{a},\mathbf{b}) \equiv \frac{\mathbf{a}^\mathsf{T}\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$.

| $\sigma(t)$ | $\mathbf{\Phi}(\mathbf{a},\mathbf{b})$ |
|---|---|
| $t$ | $\mathbf{a}^\mathsf{T}\mathbf{b}$ |
| $\max(t,0) \equiv \mathrm{ReLU}(t)$ | $\frac{1}{2\pi}\|\mathbf{a}\|\|\mathbf{b}\| \left( \angle(\mathbf{a},\mathbf{b}) \arccos\left(-\angle(\mathbf{a},\mathbf{b})\right) + \sqrt{1-\angle(\mathbf{a},\mathbf{b})^2} \right)$ |
| $\lvert t \rvert$ | $\frac{2}{\pi}\|\mathbf{a}\|\|\mathbf{b}\| \left( \angle(\mathbf{a},\mathbf{b}) \arcsin\left(\angle(\mathbf{a},\mathbf{b})\right) + \sqrt{1-\angle(\mathbf{a},\mathbf{b})^2} \right)$ |
| $\varsigma_+ \max(t,0) + \varsigma_- \max(-t,0)$ | $\frac{1}{2}(\varsigma_+^2+\varsigma_-^2)\mathbf{a}^\mathsf{T}\mathbf{b} + \frac{\|\mathbf{a}\|\|\mathbf{b}\|}{2\pi}(\varsigma_+-\varsigma_-)^2 \left( \sqrt{1-\angle(\mathbf{a},\mathbf{b})^2} - \angle(\mathbf{a},\mathbf{b}) \arccos(\angle(\mathbf{a},\mathbf{b})) \right)$ |
| $1_{t>0}$ | $\frac{1}{2} - \frac{1}{2\pi}\arccos\left(\angle(\mathbf{a},\mathbf{b})\right)$ |
| $\mathrm{sign}(t)$ | $\frac{2}{\pi}\arcsin\left(\angle(\mathbf{a},\mathbf{b})\right)$ |
| $\varsigma_2 t^2 + \varsigma_1 t + \varsigma_0$ | $\varsigma_2^2\left(2\left(\mathbf{a}^\mathsf{T}\mathbf{b}\right)^2 + \|\mathbf{a}\|^2\|\mathbf{b}\|^2\right) + \varsigma_1^2\mathbf{a}^\mathsf{T}\mathbf{b} + \varsigma_2\varsigma_0\left(\|\mathbf{a}\|^2+\|\mathbf{b}\|^2\right) + \varsigma_0^2$ |
| $\cos(t)$ | $\exp\left(-\frac{1}{2}\left(\|\mathbf{a}\|^2+\|\mathbf{b}\|^2\right)\right)\cosh(\mathbf{a}^\mathsf{T}\mathbf{b})$ |
| $\sin(t)$ | $\exp\left(-\frac{1}{2}\left(\|\mathbf{a}\|^2+\|\mathbf{b}\|^2\right)\right)\sinh(\mathbf{a}^\mathsf{T}\mathbf{b})$ |
| $\mathrm{erf}(t)$ | $\frac{2}{\pi}\arcsin\left(\frac{2\mathbf{a}^\mathsf{T}\mathbf{b}}{\sqrt{(1+2\|\mathbf{a}\|^2)(1+2\|\mathbf{b}\|^2)}}\right)$ |
| $\exp(-\frac{t^2}{2})$ | $\frac{1}{\sqrt{(1+\|\mathbf{a}\|^2)(1+\|\mathbf{b}\|^2)-(\mathbf{a}^\mathsf{T}\mathbf{b})^2}}$ |

Let us now introduce the key steps of our present analysis. Under Assumption 1, observe that for $\mathbf{x}_i \in \mathcal{C}_a$ and $\mathbf{x}_j \in \mathcal{C}_b$, $i \neq j$,

$$\mathbf{x}_i^\mathsf{T}\mathbf{x}_j = \underbrace{\boldsymbol{\omega}_i^\mathsf{T}\boldsymbol{\omega}_j}_{O(p^{-1/2})} + \underbrace{\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\mu}_b/p + \boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\omega}_j/\sqrt{p} + \boldsymbol{\mu}_b^\mathsf{T}\boldsymbol{\omega}_i/\sqrt{p}}_{O(p^{-1})}$$

which allows one to perform a Taylor expansion around 0 as $p, T \to \infty$, to give a reasonable approximation of nonlinear functions of $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$, such as those appearing in $\mathbf{\Phi}_{ij}$ (see again Table 1). For $i = j$, one has instead

$$\|\mathbf{x}_i\|^2 = \underbrace{\|\boldsymbol{\omega}_i\|^2}_{O(1)} + \underbrace{\|\boldsymbol{\mu}_a\|^2/p + 2\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\omega}_i/\sqrt{p}}_{O(n^{-1})}.$$

From $\mathbb{E}_{\boldsymbol{\omega}_i}[\|\boldsymbol{\omega}_i\|^2] = \mathrm{tr}(\mathbf{C}_a)/p$ it is convenient to further write $\|\boldsymbol{\omega}_i\|^2 = \mathrm{tr}(\mathbf{C}_a)/p + \left(\|\omega_i\|^2 - \mathrm{tr}(\mathbf{C}_a)/p\right)$, where $\mathrm{tr}(\mathbf{C}_a)/p = O(1)$ and $\|\omega_i\|^2 - \mathrm{tr}(\mathbf{C}_a)/p = O(n^{-1/2})$. By definition $\tau \equiv \mathrm{tr}(\mathbf{C}^\circ)/p = O(1)$ and exploiting again Assumption 1 one results in,

$$\|\mathbf{x}_i\|^2 = \underbrace{\tau}_{O(1)} + \underbrace{\mathrm{tr}(\mathbf{C}_a^\circ)/p + \|\boldsymbol{\omega}_i\|^2 - \mathrm{tr}(\mathbf{C}_a)/p}_{O(n^{-1/2})} + \underbrace{\|\boldsymbol{\mu}_a\|^2/p + 2\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\omega}_i/\sqrt{p}}_{O(n^{-1})}$$

which allows for a Taylor expansion of nonlinear functions of $\|\mathbf{x}_i\|^2$ around $\tau$, as has been done for $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$.

From Table 1, it appears that, for every listed $\sigma(\cdot)$, $\mathbf{\Phi}(\mathbf{x}_i,\mathbf{x}_j)$ is a smooth function of $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$ and $\|\mathbf{x}_i\|$, $\|\mathbf{x}_j\|$, despite their possible discontinuities (for example, the $\mathrm{ReLU}$ function and $\sigma(t) = |t|$). The above results therefore allow for an entry-wise Taylor expansion of the matrix $\mathbf{\Phi}$ in the large $p, T$ limit.

A critical aspect of the analysis where random matrix theory comes into play now consists in developing $\mathbf{\Phi}$ as a sum of matrices arising from the Taylor expansion and ignoring terms that give rise to a vanishing operator norm, so as to find an asymptotic equivalent matrix $\tilde{\mathbf{\Phi}}$ such that $\|\mathbf{\Phi} - \tilde{\mathbf{\Phi}}\| \to 0$ as $p, T \to \infty$, as described in detail in the following section. This analysis provides a simplified asymptotically equivalent expression for $\mathbf{\Phi}$ with all nonlinearities removed, which is the crux of the present study.

## 3. Main Results

In the remainder of this article, we shall use the following notations for random elements,

$$\mathbf{\Omega} \equiv \left[\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_T\right] \in \mathbb{R}^{p\times T}$$
$$\boldsymbol{\phi} \equiv \left\{\|\boldsymbol{\omega}_i\|^2 - \mathbb{E}\left[\|\boldsymbol{\omega}_i\|^2\right]\right\}_{i=1}^T \in \mathbb{R}^T$$

as well as for deterministic elements[2],

$$\mathbf{M} \equiv \left[\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\right] \in \mathbb{R}^{p\times K}$$
$$\mathbf{t} \equiv \left\{\frac{1}{\sqrt{p}}\,\mathrm{tr}\,\mathbf{C}_a^\circ\right\}_{a=1}^K \in \mathbb{R}^K$$

[2] As a reminder here, $\mathbf{M}$ stands for *means*, $\mathbf{t}$ accounts for (difference in) *traces* while $\mathbf{S}$ for the "*shapes*" of covariances.

$$\mathbf{J} \equiv \left[\mathbf{j}_1, \ldots, \mathbf{j}_K\right] \in \mathbb{R}^{T \times K}$$

$$\mathbf{S} \equiv \left\{ \frac{1}{p} \operatorname{tr}(\mathbf{C}_a \mathbf{C}_b) \right\}_{a,b=1}^{K} \in \mathbb{R}^{K \times K}$$

where $\mathbf{j}_a \in \mathbb{R}^T$ denotes the canonical vector of class $\mathcal{C}_a$ such that $(\mathbf{j}_a)_i = \delta_{\mathbf{x}_i \in \mathcal{C}_a}$.

**Theorem 1** (Asymptotic Equivalent of $\Phi_c$)**.** *Let Assumption 1 hold and $\mathbf{\Phi}_c$ be defined as $\mathbf{\Phi}_c \equiv \mathbf{P}\mathbf{\Phi}\mathbf{P}$, with $\mathbf{\Phi}$ given in (1). Then, as $T \to \infty$, for all $\sigma(\cdot)$ given in Table 1,*

$$\|\mathbf{\Phi}_c - \tilde{\mathbf{\Phi}}_c\| \to 0$$

*almost surely, with $\tilde{\mathbf{\Phi}}_c = \mathbf{P}\tilde{\mathbf{\Phi}}\mathbf{P}$ and*

$$\tilde{\mathbf{\Phi}} \equiv d_1 \left( \mathbf{\Omega} + \mathbf{M}\frac{\mathbf{J}^\mathsf{T}}{\sqrt{p}} \right)^\mathsf{T} \left( \mathbf{\Omega} + \mathbf{M}\frac{\mathbf{J}^\mathsf{T}}{\sqrt{p}} \right) + d_2 \mathbf{U}\mathbf{B}\mathbf{U}^\mathsf{T} + d_0 \mathbf{I}_T$$

*where we recall that $\mathbf{P} \equiv \mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^\mathsf{T}$ and*

$$\mathbf{U} \equiv \left[ \tfrac{\mathbf{J}}{\sqrt{p}}, \boldsymbol{\phi} \right]$$

$$\mathbf{B} \equiv \begin{bmatrix} \mathbf{t}\mathbf{t}^\mathsf{T} + 2\mathbf{S} & \mathbf{t} \\ \mathbf{t}^\mathsf{T} & 1 \end{bmatrix}$$

*with the coefficients $d_0, d_1, d_2$ given in Table 2.*

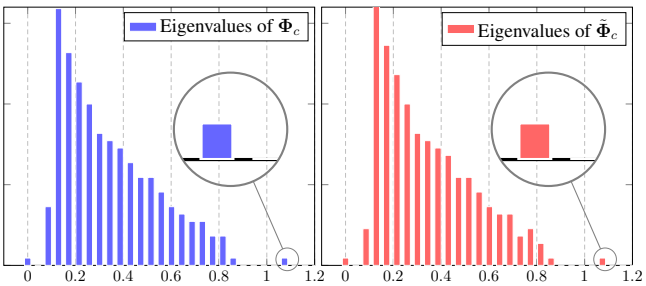


*Figure 1.* Eigenvalue distribution of $\mathbf{\Phi}_c$ and $\tilde{\mathbf{\Phi}}_c$ for the ReLU function and Gaussian mixture data with $\boldsymbol{\mu}_a = \left[\mathbf{0}_{a-1}; 3; \mathbf{0}_{p-a}\right]$, $\mathbf{C}_a = \left(1 + \frac{2(a-1)}{\sqrt{p}}\right)\mathbf{I}_p$, $p = 512$, $T = 256$ and $c_1 = c_2 = \frac{1}{2}$. Expectation estimated by averaging over 500 realizations of $\mathbf{W}$.

Theorem 1 tells us as a corollary (from Corollary 4.3.15 in (Horn & Johnson, 2012), for example) that the maximal difference between the eigenvalues of $\mathbf{\Phi}_c$ and $\tilde{\mathbf{\Phi}}_c$ vanishes asymptotically as $p, T \to \infty$, as confirmed in Figure 1. Similarly the distance between the "isolated eigenvectors[3]" also vanishes, as seen in Figure 2. This is of tremendous importance as the determination of the leading eigenvalues and eigenvectors of $\mathbf{\Phi}_c$ (that contain crucial information for clustering, for example) can be studied from the equivalent problem performed on $\tilde{\mathbf{\Phi}}_c$ and becomes mathematically more tractable.

[3]Eigenvectors that correspond to the eigenvalues found at a non-vanishing distance from the other eigenvalues.

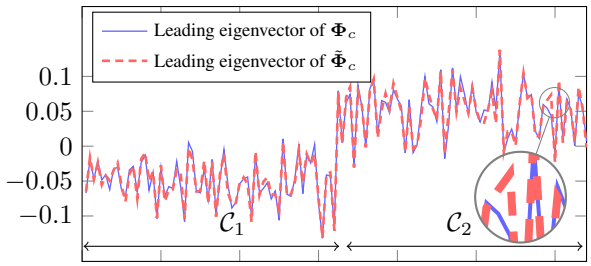


*Figure 2.* Leading eigenvector of $\mathbf{\Phi}_c$ and $\tilde{\mathbf{\Phi}}_c$ in the settings of Figure 1, with $\mathbf{j}_1 = \left[\mathbf{1}_{T_1}; \mathbf{0}_{T_2}\right]$ and $\mathbf{j}_2 = \left[\mathbf{0}_{T_1}; \mathbf{1}_{T_2}\right]$.

On closer inspection of Theorem 1, the matrix $\tilde{\mathbf{\Phi}}$ is expressed as the sum of three terms, weighted respectively by the three coefficients $d_0, d_1$ and $d_2$, that depend on the nonlinear function $\sigma(\cdot)$ via Table 2. Note that the statistical structure of the data $\{\mathbf{x}_i\}_{i=1}^T$ (namely the means in $\mathbf{M}$ and the covariances in $\mathbf{t}$ and $\mathbf{S}$) is perturbed by random fluctuations ($\mathbf{\Omega}$ and $\boldsymbol{\phi}$) and it is thus impossible to get rid of these noisy terms by wisely choosing the function $\sigma(\cdot)$. This is in sharp contrast to (Couillet et al., 2016) where it is shown that more general kernels (i.e., not arising from random feature maps) allow for a more flexible treatment of information versus noise.

However, there does exist a balance between the means and covariances, that provides some instructions in the appropriate choice of the nonlinearity. From Table 2, the functions $\sigma(\cdot)$ can be divided into the following three groups:

- *mean-oriented*, where $d_1 \neq 0$ while $d_2 = 0$: this is the case of the functions $t$, $1_{t>0}$, $\operatorname{sign}(t)$, $\sin(t)$ and $\operatorname{erf}(t)$, which asymptotically track only the difference in means (i.e., $\mathbf{t}$ and $\mathbf{S}$ disappear from the expression of $\tilde{\mathbf{\Phi}}_c$). As an illustration, in Figure 3 one fails to separate two Gaussian datasets of common mean but of different covariances with the erf function, while ReLU is able to accomplish the task;
- *covariance-oriented*, where $d_1 = 0$ while $d_2 \neq 0$: this concerns the functions $|t|$, $\cos(t)$ and $\exp(-t^2/2)$, which asymptotically track only the difference in covariances. Figure 4 illustrates the impossibility to classify Gaussian mixture with same covariance with $\sigma(t) = |t|$, in contrast to the ReLU function;
- *balanced*, where both $d_1, d_2 \neq 0$: here for the ReLU function $\max(t, 0)$, the Leaky ReLU function (Maas et al., 2013) $\varsigma_+ \max(t, 0) + \varsigma_- \max(-t, 0)$ and the quadratic function $\varsigma_2 t^2 + \varsigma_1 t + \varsigma_0$.

Before entering into a more detailed discussion of Theorem 1, first note importantly that, for practical interests, the

*Table 2.* Coefficients $d_i$ in $\tilde{\mathbf{\Phi}}_c$ for different $\sigma(\cdot)$.

| $\sigma(t)$ | $d_0$ | $d_1$ | $d_2$ |
|---|---|---|---|
| $t$ | $0$ | $1$ | $0$ |
| $\max(t,0) \equiv \mathrm{ReLU}(t)$ | $\left(\frac{1}{4}-\frac{1}{2\pi}\right)\tau$ | $\frac{1}{4}$ | $\frac{1}{8\pi\tau}$ |
| $\lvert t\rvert$ | $\left(1-\frac{2}{\pi}\right)\tau$ | $0$ | $\frac{1}{2\pi\tau}$ |
| $\varsigma_+ \max(t,0)+\varsigma_- \max(-t,0) \equiv \mathrm{LReLU}(t)$ | $\frac{\pi-2}{4\pi}(\varsigma_+ +\varsigma_-)^2\tau$ | $\frac{1}{4}(\varsigma_+ -\varsigma_-)^2$ | $\frac{1}{8\tau\pi}(\varsigma_+ +\varsigma_-)^2$ |
| $1_{t>0}$ | $\frac{1}{4}-\frac{1}{2\pi}$ | $\frac{1}{2\pi\tau}$ | $0$ |
| $\mathrm{sign}(t)$ | $1-\frac{2}{\pi}$ | $\frac{2}{\pi\tau}$ | $0$ |
| $\varsigma_2 t^2+\varsigma_1 t+\varsigma_0$ | $2\tau^2\varsigma_2^2$ | $\varsigma_1^2$ | $\varsigma_2^2$ |
| $\cos(t)$ | $\frac{1}{2}+\frac{e^{-2\tau}}{2}-e^{-\tau}$ | $0$ | $\frac{e^{-\tau}}{4}$ |
| $\sin(t)$ | $\frac{1}{2}-\frac{e^{-2\tau}}{2}-\tau e^{-\tau}$ | $e^{-\tau}$ | $0$ |
| $\mathrm{erf}(t)$ | $\frac{2}{\pi}\left(\arccos\left(\frac{2\tau}{2\tau+1}\right)-\frac{2\tau}{2\tau+1}\right)$ | $\frac{4}{\pi}\frac{1}{2\tau+1}$ | $0$ |
| $\exp(-\frac{t^2}{2})$ | $\frac{1}{\sqrt{2\tau+1}}-\frac{1}{\tau+1}$ | $0$ | $\frac{1}{4(\tau+1)^3}$ |

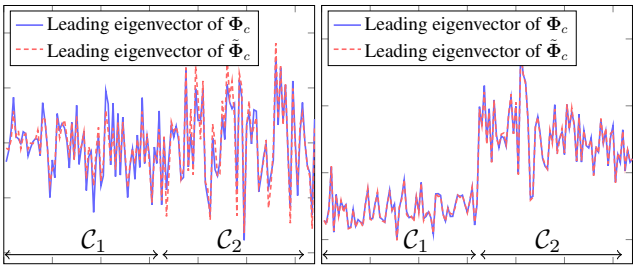


*Figure 3.* Leading eigenvector of $\mathbf{\Phi}_c$ and $\tilde{\mathbf{\Phi}}_c$ for erf (left) and the ReLU (right) function, performed on Gaussian mixture data with $\boldsymbol{\mu}_a = \mathbf{0}_p$, $\mathbf{C}_a = \left(1+\frac{15(a-1)}{\sqrt{p}}\right)\mathbf{I}_p$, $p = 512$, $T = 256$, $c_1 = c_2 = \frac{1}{2}$; $\mathbf{j}_1 = \left[\mathbf{1}_{T_1}; \mathbf{0}_{T_2}\right]$ and $\mathbf{j}_2 = \left[\mathbf{0}_{T_1}; \mathbf{1}_{T_2}\right]$. Expectation estimated by averaging over 500 realizations of $\mathbf{W}$.

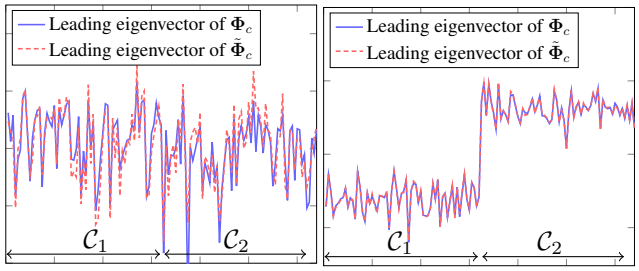


*Figure 4.* Leading eigenvector of $\mathbf{\Phi}_c$ and $\tilde{\mathbf{\Phi}}_c$ for $\sigma(t) = |t|$ (left) and the ReLU (right) function, performed on Gaussian mixture data with $\boldsymbol{\mu}_a = \left[\mathbf{0}_{a-1}; 5; \mathbf{0}_{p-a}\right]$, $\mathbf{C}_a = \mathbf{I}_p$, $p = 512$, $T = 256$, $c_1 = c_2 = \frac{1}{2}$ and $\mathbf{j}_1 = \left[\mathbf{1}_{T_1}; \mathbf{0}_{T_2}\right]$ and $\mathbf{j}_2 = \left[\mathbf{0}_{T_1}; \mathbf{1}_{T_2}\right]$. Expectation estimated by averaging over 500 realizations of $\mathbf{W}$.

quantity $\tau$ can be estimated consistently from the data, as described in the following lemma.

**Lemma 1** (Consistent estimator of $\tau$)**.** *Let Assumption 1 hold and recall the definition* $\tau \equiv \frac{1}{p}\mathrm{tr}\,(\mathbf{C}^\circ)$*. Then, as* $T \to \infty$*, with probability* 1

$$\frac{1}{T}\sum_{i=1}^{T}\|\mathbf{x}_i\|^2 - \tau \to 0.$$

*Proof.* Since

$$\frac{1}{T}\sum_{i=1}^{T}\|\mathbf{x}_i\|^2 = \frac{1}{T}\sum_{a=1}^{K}\sum_{i=1}^{T_a}\frac{1}{p}\|\boldsymbol{\mu}_a\|^2 - \frac{2}{\sqrt{p}}\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\omega}_i + \|\boldsymbol{\omega}_i\|^2,$$

with Assumption 1 we have $\frac{1}{T}\sum_{a=1}^{K}\sum_{i=1}^{T_a}\frac{1}{p}\|\boldsymbol{\mu}_a\|^2 = O(\frac{1}{p})$. The term $\frac{1}{T}\sum_{a=1}^{K}\sum_{i=1}^{T_a}\frac{2}{\sqrt{p}}\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\omega}_i$ is a linear combination of independent zero-mean Gaussian variables and vanishes with probability 1 as $p, T \to \infty$ with Chebyshev's inequality and the Borel-Cantelli lemma. Ultimately by the strong law of large numbers, we have $\frac{1}{T}\sum_{i=1}^{T}\|\boldsymbol{\omega}_i\|^2 - \tau \to 0$ almost surely, which concludes the proof. □

From a practical aspect, a few remarks on the conclusions of Theorem 1 can be made.

**Remark 1** (Constant shift in feature space)**.** *For* $\sigma(t) = \varsigma_2 t^2 + \varsigma_1 t + \varsigma_0$*, note the absence of* $\varsigma_0$ *in Table 2, meaning that the constant of the quadratic function does not affect the spectrum of the feature matrix. This phenomenon is in fact universal as it can be shown through the integral trick of (Williams, 1997) that the function* $\sigma(t) + c$ *for some constant shift* $c$ *gives the same matrix* $\mathbf{\Phi}_c$ *as the original function* $\sigma(t)$*.*

A direct consequence of Remark 1 is that the coefficients

$d_0, d_1, d_2$ of the function $\text{sign}(t)$ are four times those of $1_{t>0}$, as a result of the fact that $\text{sign}(t) = 2 \cdot 1_{t>0} - 1$. Constant shifts have, as such, no consequence in classification applications.

**Remark 2** (Universality of quadratic and Leaky ReLU functions)**.** *Ignoring the coefficient $d_0$ that gives rise to a constant shift of all eigenvalues of $\tilde{\mathbf{\Phi}}_c$ and thus of no practical relevance, observe from Table 2 that by tuning the parameters of the quadratic and Leaky* ReLU *functions (*LReLU$(t)$*), one can select arbitrary positive value for the ratio $d_1/d_2$, while the other listed functions have constraints linking $d_1$ to $d_2$.*

Following the discussions in Remark 2, the parameters $\varsigma_+, \varsigma_-$ of the LReLU, as well as $\varsigma_1, \varsigma_2$ of the quadratic function, essentially act to balance the weights of means and covariances in the mixture model of the data. More precisely, as $\frac{\varsigma_+}{\varsigma_-} \to 1$ or $\varsigma_2 \gg \varsigma_1$, more emphasis is set on the "distance" between covariance matrices while $\frac{\varsigma_+}{\varsigma_-} \to -1$ or $\varsigma_1 \gg \varsigma_2$ stresses the differences in means.

In Figure 5, spectral clustering on four classes of Gaussian data is performed: $\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1)$, $\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_2)$, $\mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_2)$ with the LReLU function that takes different values for $\varsigma_+$ and $\varsigma_-$. For $a = 1, 2$, $\boldsymbol{\mu}_a = [\mathbf{0}_{a-1}; 5; \mathbf{0}_{p-a}]$ and $\mathbf{C}_a = \left(1 + \frac{15(a-1)}{\sqrt{p}}\right) \mathbf{I}_p$. By choosing $\varsigma_+ = \varsigma_- = 1$ (equivalent to the linear map $\sigma(t) = t$) and $\varsigma_+ = -\varsigma_- = 1$ (equivalent to $\sigma(t) = |t|$), with the leading two eigenvectors we always recover two classes instead of four, as each setting of parameters only allows for a part of the statistical information of the data to be used for clustering. However, by taking $\varsigma_+ = 1, \varsigma_- = 0$ (the ReLU function) we distinguish all four classes in the leading two eigenvectors, to which the k-means method can then be applied for final classification, as shown in Figure 6.

Of utmost importance for random feature-based spectral methods (such as kernel spectral clustering discussed above (Ng et al., 2002)) is the presence of informative eigenvectors in the spectrum of $\mathbf{G}$, and thus of $\mathbf{\Phi}_c$. To gain a deeper understanding on the spectrum of $\mathbf{\Phi}_c$, one can rewrite $\tilde{\mathbf{\Phi}}$ in the more compact form,

$$\tilde{\mathbf{\Phi}} = d_1 \mathbf{\Omega}^\mathsf{T} \mathbf{\Omega} + \mathbf{V} \mathbf{A} \mathbf{V}^\mathsf{T} + d_0 \mathbf{I}_T$$

where

$$\mathbf{V} \equiv \left[ \tfrac{\mathbf{J}}{\sqrt{p}}, \boldsymbol{\phi}, \mathbf{\Omega}^\mathsf{T} \mathbf{M} \right]$$

$$\mathbf{A} \equiv \begin{bmatrix} \mathbf{A}_{11} & d_2 \mathbf{t} & d_1 \mathbf{I}_K \\ d_2 \mathbf{t}^\mathsf{T} & d_2 & 0 \\ d_1 \mathbf{I}_K & 0 & 0 \end{bmatrix}$$

with

$$\mathbf{A}_{11} \equiv d_1 \mathbf{M}^\mathsf{T} \mathbf{M} + d_2 (\mathbf{t}\mathbf{t}^\mathsf{T} + 2\mathbf{S})$$

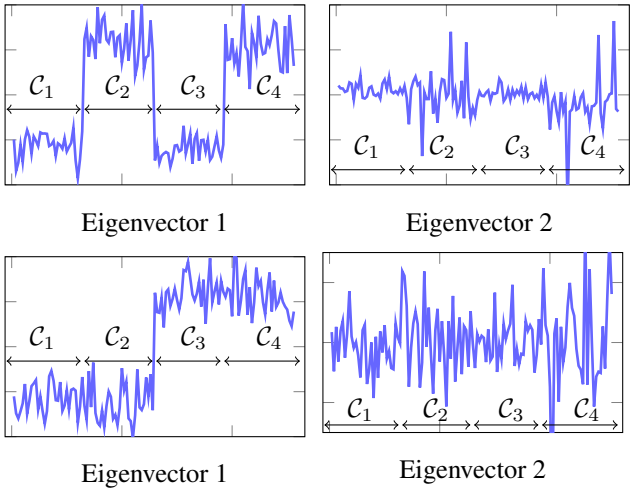


*Figure 5.* Leading two eigenvectors of $\mathbf{\Phi}_c$ for the LReLU function with $\varsigma_+ = \varsigma_- = 1$ (top) and $\varsigma_+ = -\varsigma_- = 1$ (bottom), performed on four classes Gaussian mixture data with $p = 512$, $T = 256$, $c_a = \frac{1}{4}$ and $\mathbf{j}_a = [\mathbf{0}_{T_{a-1}}; \mathbf{1}_{T_a}; \mathbf{0}_{T-T_a}]$, for $a = 1, 2, 3, 4$. Expectation estimated by averaging over 500 realizations of $\mathbf{W}$.

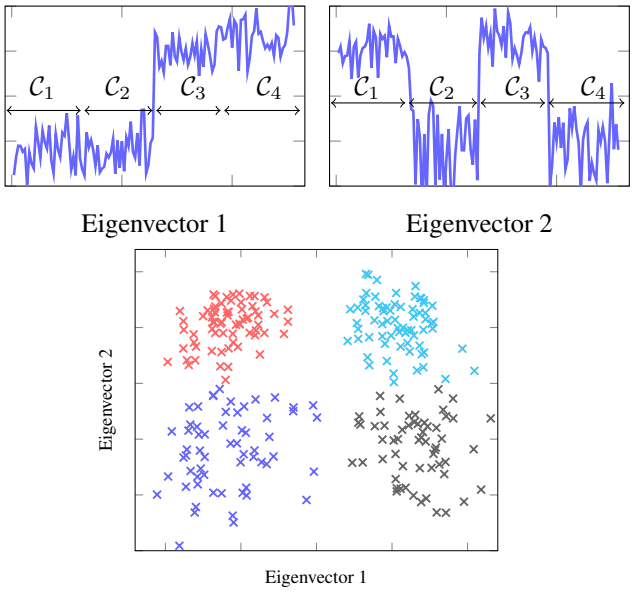


*Figure 6.* Leading two eigenvectors of $\mathbf{\Phi}_c$ (top) for the LReLU function with $\varsigma_+ = 1, \varsigma_- = 0$ and two dimensional representation of these eigenvectors (bottom), in the same setting as in Figure 5.

which is akin to the so-called "spiked model" in the random matrix literature (Baik et al., 2005), as it equals, if $d_1 \neq 0$, the sum of some standard (noise-like) random matrix $\mathbf{\Omega}^\mathsf{T} \mathbf{\Omega}$, and a low rank (here up to $2K + 1$) informative matrix $\mathbf{V} \mathbf{A} \mathbf{V}^\mathsf{T}$, that may induce some isolated eigenvalues outside the main bulk of eigenvalues in the spectrum of $\tilde{\mathbf{\Phi}}_c$, as shown in Figure 1.

The eigenvectors associated to these eigenvalues often contain crucial information about the data statistics (the classes in a classification settings). In particular, note that the matrix $\mathbf{V}$ contains the canonical vector $\mathbf{j}_a$ of class $\mathcal{C}_a$ and we thus hope to find some isolated eigenvector of $\mathbf{\Phi}_c$ aligned to $\mathbf{j}_a$ that can be directly used to perform clustering. Intuitively speaking, if the matrix $\mathbf{A}$ contains sufficient energy (has sufficiently large operator norm), the eigenvalues associated to the small rank matrix $\mathbf{VAV}^\mathsf{T}$ may jump out from the main bulk of $\mathbf{\Omega}^\mathsf{T}\mathbf{\Omega}$ and becomes "isolated" as in Figure 1, referred to as the *phase transition* phenomenon in the random matrix literature (Baik et al., 2005). The associated eigenvectors then tend to align to linear combinations of the canonical vectors $\mathbf{j}_a$ as seen in Figure 5-6. This alignment between the isolated eigenvectors and $\mathbf{j}_a$ is essentially measured by the amplitude of the eigenvalues of the matrix $\mathbf{A}_{11}$, or more concretely, the statistical differences of the data (namely, $\mathbf{t}$, $\mathbf{S}$ and $\mathbf{M}$). Therefore, a good adaptation of the ratio $d_1/d_2$ ensures the (asymptotic) detectability of different classes from the spectrum of $\mathbf{\Phi}_c$.

## 4. Numerical Validations

We complete this article by showing that our theoretical results, derived from Gaussian mixture models, show an unexpected close match in practice when applied to some real-world datasets. We consider two different types of classification tasks: one on handwritten digits of the popular MNIST (LeCun et al., 1998) database (number 6 and 8), and the other on epileptic EEG time series data (Andrzejak et al., 2001) (set B and E). These two datasets are typical examples of means-dominant (handwritten digits recognition) and covariances-dominant (EEG times series classification) tasks. This is numerically confirmed in Table 3.

*Table 3.* Empirical estimation of (normalized) differences in means and covariances of the MNIST (Figure 7 and 8) and epileptic EEG (Figure 9 and 10) datasets.

| | $\|\mathbf{M}^\mathsf{T}\mathbf{M}\|$ | $\|\mathbf{t}\mathbf{t}^\mathsf{T} + 2\mathbf{S}\|$ |
|---|---|---|
| MNIST DATA | 172.4 | 86.0 |
| EEG DATA | 1.2 | 182.7 |

### 4.1. Handwritten digits recognition

We perform random feature-based spectral clustering on data matrices that consist of $T = 32, 64$ and 128 randomly selected vectorized images of size $p = 784$ from the MNIST dataset. Means and covariances are empirically obtained from the full set of 11 769 MNIST images (5 918 images of number 6 and 5 851 of number 8). Comparing the matrix $\mathbf{\Phi}_c$ built from the data and the theoretically equivalent $\tilde{\mathbf{\Phi}}_c$ obtained as if the data were Gaussian with the (empirically) computed means and covariances, we observe an extremely close fit in the behavior of the eigenvalues in Figure 7, as well of the leading eigenvector in Figure 8. The k-means method is then applied to the leading two eigenvectors of the matrix $\mathbf{G}_c$ that consists of $n = 32$ random features to perform unsupervised classification, with resulting accuracies (averaged over 50 runs) reported in Table 4. As remarked from Table 3, the mean-oriented $\sigma(t)$ functions are expected to outperform the covariance-oriented ones in this task, which is consistent with the results in Table 4.

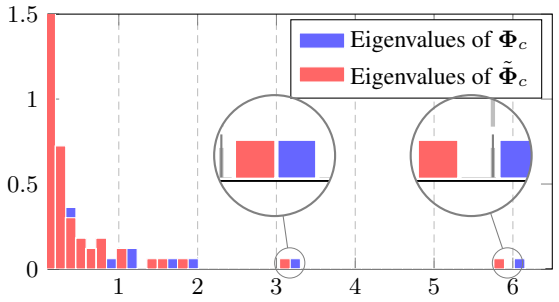


*Figure 7.* Eigenvalue distribution of $\mathbf{\Phi}_c$ and $\tilde{\mathbf{\Phi}}_c$ for the MNIST data, with the ReLU function, $p = 784$, $T = 128$ and $c_1 = c_2 = \frac{1}{2}$, with $\mathbf{j}_1 = [\mathbf{1}_{T_1}; \mathbf{0}_{T_2}]$ and $\mathbf{j}_2 = [\mathbf{0}_{T_1}; \mathbf{1}_{T_2}]$. Expectation estimated by averaging over 500 realizations of $\mathbf{W}$.

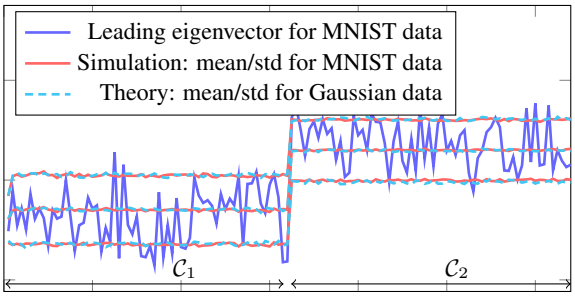


*Figure 8.* Leading eigenvector of $\mathbf{\Phi}_c$ for the MNIST and Gaussian mixture data with a width of $\pm 1$ standard deviations (generated from 500 trials) in the settings of Figure 7.

### 4.2. EEG time series classification

The epileptic EEG dataset[4], developed by the University of Bonn, Germany, is described in (Andrzejak et al., 2001). The dataset consists of five subsets (denoted A-E), each containing 100 single-channel EEG segments of 23.6-sec duration. Sets A and B were collected from surface EEG recordings of five healthy volunteers, while sets C, D and E were collected from the EEG records of the pre-surgical diagnosis of five epileptic patients. Here we perform ran-

[4] http://www.meb.unibonn.de/epileptologie/science/physik/eegdata.html.

*Table 4.* Classification accuracies for random feature-based spectral clustering with different $\sigma(t)$ on the MNIST dataset.

| | $\sigma(t)$ | $T = 32$ | $T = 64$ | $T = 128$ |
|---|---|---|---|---|
| MEAN-ORIENTED | $t$ | 85.31% | **88.94%** | 87.30% |
| | $1_{t>0}$ | 86.00% | 82.94% | 85.56% |
| | $\text{sign}(t)$ | 81.94% | 83.34% | 85.22% |
| | $\sin(t)$ | 85.31% | 87.81% | **87.50%** |
| | $\text{erf}(t)$ | **86.50%** | 87.28% | 86.59% |
| COV-ORIENTED | $\lvert t\rvert$ | 62.81% | 60.41% | 57.81% |
| | $\cos(t)$ | 62.50% | 59.56% | 57.72% |
| | $\exp(-\frac{t^2}{2})$ | 64.00% | 60.44% | 58.67% |
| BALANCED | $\text{ReLU}(t)$ | 82.87% | 85.72% | 82.27% |

dom feature-based spectral clustering on $T = 32, 64$ and 128 randomly picked EEG segments of length $p = 100$ from the dataset. Means and covariances are empirically estimated from the full set (4 097 segments of set B and 4 097 segments of set E). Similar behavior of eigenpairs as for Gaussian mixture models is once more observed in Figure 9 and 10. After k-means classification on the leading two eigenvectors of the (centered) Gram matrix composed of $n = 32$ random features, the accuracies (averaged over 50 runs) are reported in Table 5.

As opposed to the MNIST image recognition task, from Table 5 it is easy to check that the covariance-oriented functions (i.e., $\sigma(t) = |t|$, $\cos(t)$ and $\exp(-t^2/2)$) far outperform any other with almost perfect classification accuracies. It is particularly interesting to note that the popular ReLU function is suboptimal in both tasks, but never performs very badly, thereby offering a good risk-performance tradeoff.

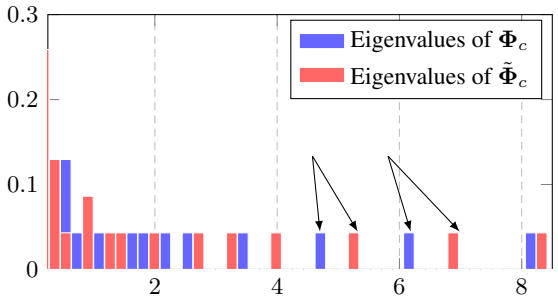


*Figure 9.* Eigenvalue distribution of $\mathbf{\Phi}_c$ and $\tilde{\mathbf{\Phi}}_c$ for the epileptic EEG, with the ReLU function, $p = 100$, $T = 128$ and $c_1 = c_2 = \frac{1}{2}$, with $\mathbf{j}_1 = \left[\mathbf{1}_{T_1}; \mathbf{0}_{T_2}\right]$ and $\mathbf{j}_2 = \left[\mathbf{0}_{T_1}; \mathbf{1}_{T_2}\right]$. Expectation estimated by averaging over 500 realizations of $\mathbf{W}$.

## 5. Conclusion

In this article, we have provided a theoretical analysis on random feature-based spectral algorithms for large dimensional data, providing a better understanding of the precise mechanism underlying these methods. Our results show a quite simple relation between the nonlinear function involved in the random feature map (only through two scalars $d_1$ and $d_2$) and the capacity of the latter to discriminate data upon their means and covariances. In obtaining this result, we demonstrated that point-wise nonlinearities can be incorporated into a classical Taylor expansion as a consequence of the concentration phenomenon in high dimensional space. This result was then validated through experimental classification tasks on the MNIST and EEG datasets.

This paper can be taken as a first step of the random matrix-based understanding and improvements of various learning methods using random features, for example the so-called extreme learning machine (Huang et al., 2012), essentially being a ridge regression on random features; as well as of more elaborate neural networks (Lillicrap et al., 2016), the nonlinear activation of which being the main difficulty for a thorough analysis. Moreover, following recent advances in random matrix analysis of learning methods (Ali & Couillet, 2016), it is envisioned to estimate consistently the hyperparameter $d_1/d_2$ of utmost importance and thus improve the performance of all random feature-based methods.

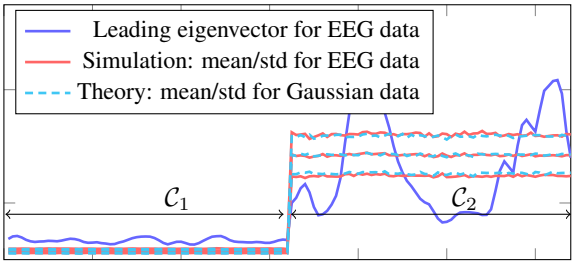


*Figure 10.* Leading eigenvector of $\mathbf{\Phi}_c$ for the EEG and Gaussian mixture data with a width of $\pm 1$ standard deviation (generated from 500 trials) in the settings of Figure 9.

*Table 5.* Classification accuracies for random feature-based spectral clustering with different $\sigma(t)$ on the epileptic EEG dataset.

| | $\sigma(t)$ | $T = 32$ | $T = 64$ | $T = 128$ |
|---|---|---|---|---|
| MEAN-ORIENTED | $t$ | 71.81% | 70.31% | 69.58% |
| | $1_{t>0}$ | 65.19% | 65.87% | 63.47% |
| | $\text{sign}(t)$ | 67.13% | 64.63% | 63.03% |
| | $\sin(t)$ | 71.94% | 70.34% | 68.22% |
| | $\text{erf}(t)$ | 69.44% | 70.59% | 67.70% |
| COV-ORIENTED | $\lvert t\rvert$ | 99.69% | 99.69% | 99.50% |
| | $\cos(t)$ | 99.00% | 99.38% | 99.36% |
| | $\exp(-\frac{t^2}{2})$ | **99.81%** | **99.81%** | **99.77%** |
| BALANCED | $\text{ReLU}(t)$ | 84.50% | 87.91% | 90.97% |

## References


Ali, Hafiz Tiomoko and Couillet, Romain. Spectral community detection in heterogeneous large networks. *arXiv preprint arXiv:1611.01096*, 2016.

Andrzejak, Ralph G, Lehnertz, Klaus, Mormann, Florian, Rieke, Christoph, David, Peter, and Elger, Christian E. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6):061907, 2001.

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

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

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

Horn, Roger A and Johnson, Charles R. *Matrix analysis*. Cambridge university press, 2012.

Huang, Guang-Bin, Zhou, Hongming, Ding, Xiaojian, and Zhang, Rui. Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(2):513–529, 2012.

Keriven, Nicolas, Bourrier, Anthony, Gribonval, Rémi, and Pérez, Patrick. Sketching for large-scale learning of mixture models. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 6190–6194. IEEE, 2016.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

LeCun, Yann, Cortes, Corinna, and Burges, Christopher JC. The MNIST database of handwritten digits, 1998.

Lillicrap, Timothy P, Cownden, Daniel, Tweed, Douglas B, and Akerman, Colin J. Random synaptic feedback weights support error backpropagation for deep learning. *Nature communications*, 7, 2016.

Louart, Cosme, Liao, Zhenyu, and Couillet, Romain. A random matrix approach to neural networks. *arXiv preprint arXiv:1702.05419*, 2017.

Maas, Andrew L, Hannun, Awni Y, and Ng, Andrew Y. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, volume 30, 2013.

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

Pennington, Jeffrey and Worah, Pratik. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, pp. 2634–2643, 2017.

Rahimi, Ali and Recht, Benjamin. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pp. 1177–1184, 2008.

Scardapane, Simone and Wang, Dianhui. Randomness in neural networks: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(2), 2017.

Schmidhuber, Jürgen. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

Schölkopf, Bernhard and Smola, Alexander J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Silverstein, Jack W and Bai, ZD. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate analysis*, 54(2): 175–192, 1995.

Vedaldi, Andrea and Zisserman, Andrew. Efficient additive kernels via explicit feature maps. *IEEE transactions on pattern analysis and machine intelligence*, 34(3):480–492, 2012.

Williams, Christopher KI. Computing with infinite networks. *Advances in neural information processing systems*, pp. 295–301, 1997.