# Kernel random matrices of large concentrated data: the example of GAN-generated images

Mohamed El Amine Seddik, Mohamed Tamaazousti, Romain Couillet

# KERNEL RANDOM MATRICES OF LARGE CONCENTRATED DATA: THE EXAMPLE OF GAN-GENERATED IMAGES

*Mohamed El Amine Seddik*[1,2]*, Mohamed Tamaazousti*[1]*, Romain Couillet*[2,3]

[1]CEA, LIST, 8 Avenue de la Vauve, 91120 Palaiseau, France, [2]CentraleSupélec and [3]GIPSA-lab.

## ABSTRACT

Based on recent random matrix advances in the analysis of kernel methods for classification and clustering, this paper proposes the study of large kernel methods for a wide class of random inputs, *i.e.*, concentrated data, which are more generic than Gaussian mixtures. The concentration assumption is motivated by the fact that one can use generative models to design complex data structures, through *Lipschitz-ally* transformed concentrated vectors (*e.g.*, Gaussian) which remain concentrated vectors. Applied to spectral clustering, we demonstrate that our theoretical findings closely match the behavior of large kernel matrices, when considering the fed-in data as CNN representations of GAN-generated images (*i.e.*, concentrated vectors by design).

***Index Terms***— Kernel methods, spectral clustering, random matrix theory, concentration of measure, GANs.

## 1. INTRODUCTION

The big data paradigm involves the ability of performing classification or regression tasks on large dimensional and numerous datasets (*i.e.*, the so-called "large $p$", "large $n$" regime of random matrix theory). Generally, the used methods for achieving these tasks are based on non-linear approaches including neural networks [1, 2] and algorithms that are based on kernel methods, such as kernel-based support vector machines [3], semi-supervised classification [4], kernel-based PCA [5] and spectral clustering [6, 7]. Due to their non-linear design, these methods are particularly difficult to analyze theoretically. For practical large and numerous data, the study of kernel-based methods relies on the characterization of kernel matrices $\mathbf{K} \in \mathbb{R}^{n\times n}$ in the large dimensional regime (*i.e.*, $p/n \to c_0$ as $n \to \infty$). Under *asymptotically non-trivial* growth rate assumptions on the data statistics (*i.e.*, maintaining a feasible get not too easy problem), the entries $K_{ij} = f(\mathbf{x}_i^\intercal \mathbf{x}_j/p)$ or $K_{ij} = f(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)$ of $\mathbf{K}$ tend to a limiting constant independently of the data classes – *the between and within class vectors are "equidistant" in high-dimension*. This observation allows one to study $\mathbf{K}$ through a Taylor expansion, thereby giving access to the characterization of functionals of $\mathbf{K}$ or its (informative) eigenvalue-eigenvector pairs in the large dimensional regime.

Indeed, such an analysis was initiated in [8] where it has been shown that $\mathbf{K}$ has a linear behavior in the large $p, n$ asymptotics. Under a $k$-class Gaussian mixture model, it has been shown in [7] that the normalized Laplacian matrix associated with $\mathbf{K}$ behaves asymptotically as a so-called spiked random matrix, where some of the isolated eigenvalues and eigenvectors contain clustering information. In particular, the authors in [7] demonstrated that the obtained theoretical model agrees with empirical results using the popular MNIST dataset [9], thereby suggesting a sort of *universality* of spectral clustering regarding the underlying data distribution.

The aim of this paper is to confirm this observation by relaxing the Gaussianity assumption to a wide range of distributions. In fact, most of real world data (*e.g.*, images or CNN representations that are commonly used in computer vision [10]) belong to complex manifolds, and therefore are unlikely close to Gaussian. However, due to recent advances in generative models since the arrival of Generative Adversarial Networks [11], it is now possible to generate complex data structures by applying successive Lipschitz operations to Gaussian vectors. On the other hand, the concentration of measure phenomenon tells us that Lipschitz transformations of Gaussian vectors satisfy a concentration property [12, Thm 2.1.12]. Precisely, defining a *concentrated vector* $X \in E$ through the real concentration of $\mathcal{F}(X)$, for any Lipschitz function $\mathcal{F} : E \to \mathbb{R}$, defines a larger class of distributions [13]. This suggests that making the aforementioned *concentration* assumption on data is a suitable model for real world data.

In this paper, we analyze the kernel matrix $\mathbf{K}$ under a $k$-class *concentration* mixture model [13]. Precisely, we prove that $\mathbf{K}$ behaves (up to centering) asymptotically as a spiked random matrix in the large $p$ large $n$ regime, thereby generalizing the results of [7] to a broader class of distributions. We particularly confirm our theoretical findings by considering the input data as CNN representations of images generated by a GAN, where the latter is trained to fit the manifold distribution of the well-known CIFAR-10 dataset [14]. We further consider real images for comparison.

*Notation*: Vectors are denoted by boldface lowercase letters and matrices by boldface uppercase letters. The notation

Couillet's work is supported by the GSTATS UGA IDEX Datascience chair and the ANR RMT4GRAPH (ANR-14-CE28-0006).

$\|\cdot\|$ stands for the Euclidean norm for vectors and the operator norm for matrices. The vector $\mathbf{1}_n \in \mathbb{R}^n$ denotes the vector filled with ones. For an integer $k$, $[k]$ stands for the set $\{1, \ldots, k\}$. $[\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$ denotes the concatenation of the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$. Given a normed space $(E, \|\cdot\|_E)$ and a real $q$, an element $X \in E$ is said to be $q$-exponentially concentrated if for any 1-Lipschitz function $\mathcal{F} : E \to \mathbb{R}$, $\mathbb{P}\{|\mathcal{F}(X) - \mathbb{E}\mathcal{F}(X)| \geq t\} \leq Ce^{-ct^q}$ for all $t > 0$, and we shall write $X \in \mathcal{O}(e^{-\cdot^q})$ in $(E, \|\cdot\|_E)$.

**Remark 1.** *Let $X \in \mathcal{O}(e^{-\cdot^q})$ in $(E, \|\cdot\|_E)$ and $\mathcal{F}_n : E \to F$ $u_n$-Lipschitz, where $u_n$ depends on some asymptotic variable $n$. Then, the concentration property on $X$ is transferred to $\mathcal{F}_n(X)$, precisely $\mathcal{F}_n(X) \in \mathcal{O}(e^{-(\cdot/u_n)^q})$ in $(F, \|\cdot\|_F)$.*

## 2. MODEL SETTING AND ASSUMPTIONS

Consider $n$ independent random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ distributed in $k < \infty$ classes represented by $k$ distributions $\mu_1, \ldots, \mu_k$ supposedly all distinct. We consider the hypothesis of $q$-exponential concentration, meaning that there exists $q \geq 2$ such that for all $m \in \mathbb{N}$, any $\ell \in [k]$ and any family of independent vectors $\mathbf{y}_1, \ldots, \mathbf{y}_m$ following the distribution $\mu_\ell$, we have the concentration

$$[\mathbf{y}_1, \ldots, \mathbf{y}_m] \in \mathcal{O}(e^{-\cdot^q}) \text{ in } (\mathbb{R}^{p\times m}, \|\cdot\|_F). \tag{1}$$

For $\ell \in [k]$, we denote by $\mathbf{m}_\ell$ the mean of the distribution $\mu_\ell$, $\mathbf{C}_\ell$ denotes its covariance matrix and $n_\ell$ stands for the number of vectors among the $\mathbf{x}_i$'s following $\mu_\ell$. Let $\mathbf{m} \in \mathbb{R}^p$ and $\mathbf{C} \in \mathbb{R}^{p\times p}$ be respectively defined as

$$\mathbf{m} \equiv \sum_{\ell=1}^k \frac{n_\ell}{n}\mathbf{m}_\ell, \quad \mathbf{C} \equiv \sum_{\ell=1}^k \frac{n_\ell}{n}\mathbf{C}_\ell \tag{2}$$

We further denote $\bar{\mathbf{m}}_\ell \equiv \mathbf{m}_\ell - \mathbf{m}$ and $\bar{\mathbf{C}}_\ell \equiv \mathbf{C}_\ell - \mathbf{C}$.

We shall consider the following set of assumptions on the data statistics and the kernel function in the large dimensional regime, meaning that both $p$ and $n$ grow at controlled joint rate. These assumptions notably guarantee the non-triviality of spectral clustering under the considered regime.

**Assumption 1** (Growth rate)**.** *As $n \to \infty$, consider the following conditions:*

*i. For $c_0 \equiv \frac{p}{n}$; $0 < \liminf_n c_0 \leq \limsup_n c_0 < \infty$.*

*ii. For each $\ell \in [k]$, define $c_\ell \equiv \frac{n_\ell}{n}$ and $\mathbf{c} \equiv \{c_\ell\}_{\ell=1}^k$; $0 < \liminf_n c_\ell \leq \limsup_n c_\ell < \infty$.*

*iii. $\limsup_p \max_\ell \|\bar{\mathbf{m}}_\ell\| < \infty$, $\limsup_p \frac{\mathbb{E}\|\mathbf{x}_i\|}{\sqrt{p}} < \infty$.*

*iv. $\limsup_p \max_\ell \|\bar{\mathbf{C}}_\ell\| < \infty$, $\limsup_p \max_\ell \frac{\operatorname{tr}\bar{\mathbf{C}}_\ell}{\sqrt{p}} < \infty$.*

**Assumption 2** (Kernel function)**.** *Let $\tau \equiv \frac{2}{p}\operatorname{tr}\mathbf{C}$ and let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a three-times continuously differentiable function in a neighborhood of the values taken by $\tau$ and such that $\liminf_n f(\tau) > 0$.*

Without loss of generality, for each $\ell \in [k]$, we arrange the $\mathbf{x}_i$'s as $\mathbf{x}_{n_1+\cdots+n_{\ell-1}+1}, \ldots, \mathbf{x}_{n_1+\cdots+n_\ell} \sim \mu_\ell$. and define the kernel matrix $\mathbf{K}$ as the translation-invariant random matrix

$$\mathbf{K} \equiv \left\{ f\left(\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)\right\}_{i,j=1}^n. \tag{3}$$

## 3. MAIN RESULTS

Our first and fundamental result states that the between and within class vectors are "equidistant" in the high-dimensional regime. Namely, we have the following lemma under the $q$-exponential concentration hypothesis and Assumption 1.

**Lemma 1.** *Denote $\tau \equiv \frac{2}{p}\operatorname{tr}\mathbf{C}$ and let Assumption 1 hold. Then for any $\delta > 0$, we have with probability at least $1 - \delta$*

$$\max_{1\leq i\neq j\leq n}\left|\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 - \tau\right| = \mathcal{O}\left(\frac{\log(\frac{n}{\sqrt{\delta}})^{1/q}}{\sqrt{n}}\right). \tag{4}$$

From this observation, all the off-diagonal entries of the kernel matrix $\mathbf{K}$ tend to the same quantity $f(\tau)$ asymptotically. Therefore, $\mathbf{K}$ can be Taylor expanded entry-wise and we show in the following that it asymptotically has (up to centering) the same behavior as a spiked random matrix.

Before introducing this asymptotic equivalent and for subsequent use, we introduce the following quantities

$$\mathbf{M} = [\bar{\mathbf{m}}_1, \ldots, \bar{\mathbf{m}}_k] \in \mathbb{R}^{p\times k}, \quad \mathbf{t} = \left\{\frac{\operatorname{tr}\bar{\mathbf{C}}_\ell}{\sqrt{p}}\right\}_{\ell=1}^k \in \mathbb{R}^k$$

$$\mathbf{J} = [\mathbf{j}_1, \ldots, \mathbf{j}_k] \in \mathbb{R}^{n\times k}, \quad \mathbf{T} = \left\{\frac{\operatorname{tr}\bar{\mathbf{C}}_a\bar{\mathbf{C}}_b}{p}\right\}_{a,b=1}^k \in \mathbb{R}^{k\times k}$$

$$\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p\times n}, \quad \mathbf{P} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\intercal \in \mathbb{R}^{n\times n}$$

where $\mathbf{j}_\ell \in \mathbb{R}^n$ stands for the canonical vector of the class represented by $\mu_\ell$, defined by $(\mathbf{j}_\ell)_i = \delta_{\mathbf{x}_i\sim\mu_\ell}$. The vectors $\mathbf{z}_i$ are defined as $\mathbf{z}_i \equiv (\mathbf{x}_i - \bar{\mathbf{m}}_\ell)/\sqrt{p}$ for each $\ell \in [k]$.

Our main technical result states that there exists a matrix $\hat{\mathbf{K}}$ such that $\|\mathbf{PKP} - \hat{\mathbf{K}}\| \to 0$ asymptotically, where $\hat{\mathbf{K}}$ has a tractable behavior from a mathematical standpoint.

**Theorem 1** (Asymptotic Random Matrix Equivalent)**.** *Let Assumptions 1 and 2 hold and let $\hat{\mathbf{K}}$ be defined as*

$$\hat{\mathbf{K}} = -2f'(\tau)\left[\mathbf{PZ}^\intercal\mathbf{ZP} + \mathbf{UAU}^\intercal\right] + F(\tau)\mathbf{P}$$

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{I}_k & -\frac{f''(\tau)}{2f'(\tau)}\mathbf{t} \\ \mathbf{I}_k & \mathbf{0}_{k\times k} & \mathbf{0}_{k\times 1} \\ -\frac{f''(\tau)}{2f'(\tau)}\mathbf{t}^\intercal & \mathbf{0}_{1\times k} & -\frac{f''(\tau)}{2f'(\tau)} \end{bmatrix}$$

$$\mathbf{A}_{11} = \mathbf{M}^\intercal\mathbf{M} - \mathbf{\Xi} - \frac{f''(\tau)}{2f'(\tau)}[\mathbf{tt}^\intercal + 2\mathbf{T}]$$

$$\mathbf{U} = \left[\frac{\mathbf{J} - 1_n c^\intercal}{\sqrt{p}}, \mathbf{P\Phi}, \mathbf{P}\psi\right], \quad \mathbf{\Xi} = \left\{\frac{\|\bar{\mathbf{m}}_a\|^2 + \|\bar{\mathbf{m}}_b\|^2}{2}\right\}_{a,b=1}^k$$

$$\mathbf{\Phi} = \mathbf{Z}^\intercal \mathbf{M} - \{\mathbf{Z}_\ell^\intercal \bar{\mathbf{m}}_\ell \mathbf{1}_k^\intercal\}_{\ell=1}^k$$
$$\psi_i = \|\mathbf{z}_i\|^2 - \mathbb{E}\|\mathbf{z}_i\|^2 = \|\mathbf{z}_i\|^2 - \frac{1}{p}\operatorname{tr}\mathbf{C}_\ell$$
$$F(\tau) = (f(0) - f(\tau) + \tau f'(\tau)).$$

*For $\delta > 0$, there exists $C_\delta > 0$ such that for all $\gamma > 0$*

$$\|\mathbf{PKP} - \hat{\mathbf{K}}\| \leq C_\delta\, n^{-1/2+\gamma}\log(n)^\gamma \text{ with proba. } 1-\delta.$$

Theorem 1 shows that, up to centering by $\mathbf{P}$, the kernel matrix $\mathbf{K}$ has asymptotically the same behavior as $\hat{\mathbf{K}}$. In particular, the obtained approximation in operator norm implies that $\mathbf{K}$ and $\hat{\mathbf{K}}$ share the same eigenvalues (by Weyl's inequality [15, Thm 4.1]) and same *isolated* eigenvectors asymptotically. Therefore, the asymptotic spectral properties of $\mathbf{K}$ (*i.e.*, the classification performance of algorithms involving $\mathbf{K}$) may be studied through its equivalent $\hat{\mathbf{K}}$.

Indeed, note that $\hat{\mathbf{K}}$ is made of a sum of a random matrix $\mathbf{PZ}^\intercal\mathbf{ZP}$ and a maximum $(k-1)$-rank matrix containing linear combinations of the class-wise canonical vectors $\mathbf{j}_\ell$ weighted by the inner-products between class means $\mathbf{M}^\intercal\mathbf{M}$ and class covariance-products and traces (through $\mathbf{t}$ and $\mathbf{T}$). The matrix $\hat{\mathbf{K}}$ can then be identified as a so-called *spiked random matrix model* [16]. Note however that, unlike the standard spiked random matrices, the low-rank part of $\hat{\mathbf{K}}$ depends statistically on the noise part and the latter is a mixture between random matrices made of concentrated vectors [13].

Random matrix theory offers a wide range of tools to analyze such spiked models. Importantly, authors in [13] have characterized the spectrum of a sample covariance matrix of concentrated vectors, and more precisely the *bulk* of eigenvalues of the random matrix $\mathbf{PZ}^\intercal\mathbf{ZP}$. The spectrum of $\hat{\mathbf{K}}$ is then composed of a *bulk* along with up to $k-1$ isolated eigenvalues, and the associated eigenvectors are aligned with the eigenvectors in $\mathbf{U}$, therefore with linear combinations of the class canonical vectors $\mathbf{j}_1, \ldots, \mathbf{j}_k$. Consequently, characterizing the asymptotic performance of spectral clustering relies on the characterization of the isolated eigenvectors of $\hat{\mathbf{K}}$. In fact, these eigenvectors are *informative* if their associated eigenvalues are far away from the main eigenvalue *bulk*. In the following, we provide the conditions under which the *informative* eigenvalues become visible in the spectrum of $\hat{\mathbf{K}}$. Before introducing this phase transition phenomenon, let us introduce the following lemma which characterizes the spectrum of the random matrix $\mathbf{PZ}^\intercal\mathbf{ZP}$.

**Lemma 2** (Deterministic equivalent)**.** *Let Assumption 1 hold. Let $z \in \mathbb{C} \setminus \mathcal{S}$ with $\mathcal{S}$ introduced subsequently and define the resolvent matrix $\mathbf{Q}_\delta \equiv (\sum_{\ell=1}^k c_\ell \frac{\mathbf{C}_\ell}{1+\delta_\ell(z)} - z\mathbf{I}_p)^{-1}$ where $\delta_\ell(z)$ is the unique solution of the fixed point equation $\delta_\ell(z) = \frac{1}{n}\operatorname{tr}(\mathbf{C}_\ell\mathbf{Q}_\delta)$. Then the spectral distribution $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{\lambda_i}$ almost surely converges to a probability measure $\nu$ defined on a compact support $\mathcal{S}$ and having Stieltjes[1] transform the limit of $\frac{1}{p}\operatorname{tr}\mathbf{Q}_\delta$ as $p \to \infty$.*

[1]Defined as $m(z) = \int \frac{\nu(dt)}{t-z}$, for $z \in \mathbb{C}_+$.

Having Lemma 2 we can now determine the conditions under which the spikes can be visible outside the main *bulk* of $\mathbf{PZ}^\intercal\mathbf{ZP}$, and the result concerning the isolated eigenvectors. We however need the following technical assumption on the class-wise covariances to ensure that $\mathbf{PZ}^\intercal\mathbf{ZP}$ does not produce *non-informative* isolated eigenvalues.

**Assumption 3** (Spikes control)**.** *Denote $\lambda_1^\ell, \ldots, \lambda_p^\ell$ the eigenvalues of $\mathbf{C}_\ell$, for each $\ell \in [k]$. As $n \to \infty$, $\frac{1}{p}\sum_{i=1}^p \delta_{\lambda_i^\ell} \xrightarrow{\mathcal{D}} \rho_\ell$ with support $\mathcal{S}_\ell$, and $\max_{1\leq i\leq p}\operatorname{dist}(\lambda_i^\ell, \mathcal{S}_\ell) \to 0$.*

Now we can state the theorem that ensures the presence of *informative* eigenvalues in the spectrum of $\hat{\mathbf{K}}$, and gives the characterization of the corresponding isolated eigenvectors, which results from standard random matrix techniques [16].

**Theorem 2** (Spikes and isolated eigenvectors)**.** *Let Assumptions 1-3 hold and $z \in \mathbb{C} \setminus \mathcal{S}$. There exists a matrix $\mathbf{\Lambda}_z \in \mathbb{C}^{k\times k}$ and a complex-valued function $\alpha_\tau(z)$ both depending on the data statistics such that, if $\lambda^* \in \mathbb{R} \setminus \mathcal{S}$ with $\alpha_\tau(\lambda^*) \neq 0$ and $\mathbf{\Lambda}_{\lambda^*}$ having a zero eigenvalue of multiplicity $m^*$, then $\mathbf{PKP}$ produces $m^*$ spikes asymptotically close to $\rho^* = -2f'(\tau)\lambda^* + F(\tau)$. Furthermore, the eigenspace projector corresponding to the (asymptotically converging to $\rho^*$) isolated eigenvalues of $\mathbf{PKP}$ has a non-vanishing projection onto $\operatorname{span}(\mathbf{j}_1, \ldots, \mathbf{j}_k)$.*

Theorem 2 gives the conditions under which the spikes can be observed in the spectrum of $\hat{\mathbf{K}}$, and states that the corresponding eigenvectors are aligned to some extent to the class canonical vectors $\mathbf{j}_1, \ldots, \mathbf{j}_k$, which is important for spectral clustering. Note that, for lack of space, the explicit formulas of $\mathbf{\Lambda}_z$ and $\alpha_\tau(z)$ shall be given in an extended version of the present paper. We refer the reader to [7] for a detailed statement of Theorem 2 in the $k$-class Gaussian mixture model case.

## 4. APPLICATION TO GAN-GENERATED IMAGES

The $k$-class *concentration* mixture model is motivated, among other things, by the fact that data generated by GANs belong to this category of random vectors. To highlight this aspect, we evaluate our theoretical findings by considering $\mathbf{x}_1, \ldots, \mathbf{x}_n$ as CNN representations of GAN-generated images and we further consider real images for comparison.

**GAN architecture and training:** We train a conditional GAN [17] on the whole CIFAR-10 train set. Precisely, the generator $\mathcal{G}$ takes as input a Gaussian vector of dimension 100 and a one-hot-encoder vector corresponding to a given class, and outputs an image of size $32 \times 32 \times 3$. In particular, the considered architecture for $\mathcal{G}$ is a deep convolutional network [18] composed of four convolutional layers, each one followed by batch-normalization and ReLU activation except for the last layer where $\tanh$ is used as recommended by [18]. Examples of the generated images are shown in top of Fig. 1.

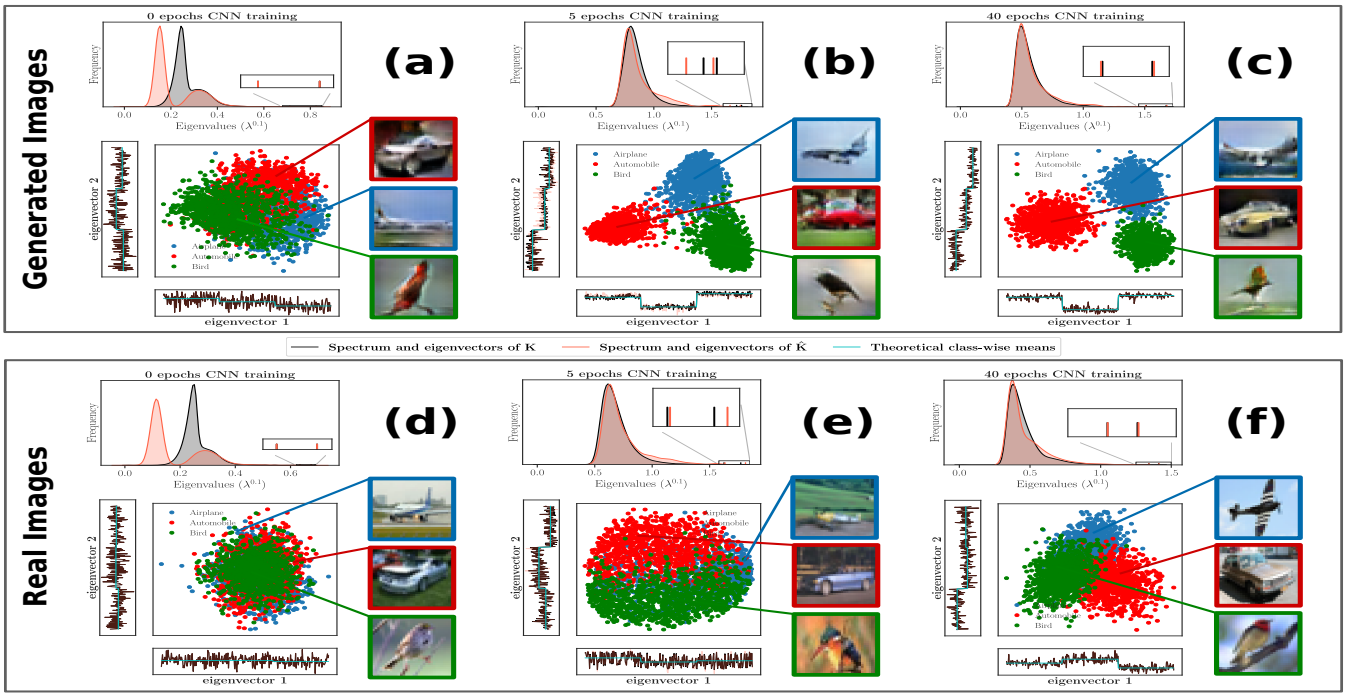


**Fig. 1**: Spectral clustering on CNN representations of GAN-generated images (top) and CIFAR-10 images (bottom) at different training phases of the representations. The performance clustering is notably predictable through random matrix theory.

**CNN representations:** In order to build CNN representations as commonly used in computer vision [10], we train two CNNs with a 10-class classification problem. A network $\mathcal{N}_g$ to classify a set of 50000 generated images (by $\mathcal{G}$), and a network $\mathcal{N}_r$ to classify the CIFAR-10 train set. The two networks have the same architecture: Six convolutional layers with ReLU activation, followed by a dense layer of 1024 units ReLU activated and a 10-units classification layer. CNN representations (denoted $\mathcal{N}_i^R$ for $i \in \{g, r\}$) correspond to the dense layer which are vectors of dimension $p = 1024$.

The $\mathbf{x}_i$'s are therefore set as $\mathbf{x}_i = \mathcal{N}_g^R \circ \mathcal{G}(\omega_i)$ for the generated images, and as $\mathbf{x}_i = \mathcal{N}_r^R(I_i)$ for the real images, where the $\omega_i$'s are random Gaussian vectors in $\mathbb{R}^{100}$ and the $I_i$'s denote images from the CIFAR-10 test set. Importantly, note that the mapping $\mathcal{N}_g^R \circ \mathcal{G}$ is Lipschitz since it is constructed from successive convolutions, activations (ReLU and $\tanh$) and batch-normalizations, which are all Lipschitz operations [19]. Therefore, by Remark 1, the $\mathbf{x}_i$'s are random *concentrated* vectors by design for the generated images.

In Fig. 1 we consider the spectral clustering of $n = 3 \times 1000$ vectors $\mathbf{x}_i$ of size $p = 1024$ belonging to $k = 3$ classes (Airplane, Automobile and Bird). In particular, we consider the clustering of these vectors at different training phases of the networks $\mathcal{N}_g$ and $\mathcal{N}_r$ (0, 5 and 40 epochs), with the generated images (top) and the CIFAR-10 test set images (bottom). Means and covariances for $\hat{\mathbf{K}}$ are computed empirically and $\mathbf{K}$ is obtained using $f(x) = \exp(-x)$.

It is important to note, from the different histograms, that the spectrum of $\mathbf{K}$ is quite close to our theoretical approximation $\hat{\mathbf{K}}$ given by Theorem 1, mainly in the cases **(b)** and **(c)** for the generated images, and even for real images in **(e)** and **(f)**. Another important aspect concerns the match between the spikes and the almost perfect match between the corresponding eigenvectors, which provide clustering information as predicted by Theorem 2. These observations notably show the *universality* aspect of spectral clustering, thereby confirming, under the concentration assumption, the observations of [7]. Beyond this result, we have shown through this paper, for the first time, that the processing of real world data (*e.g.*, GAN-generated images $\approx$ real images) can be theoretically analyzed through random matrix theory, this being made possible thanks to the concentration of measure phenomenon and the Lipschitz character of recent generative methods and their strong performances to generate complex data structures.

## 5. CONCLUSION

In this paper, we have analyzed large kernel matrices under a $k$-class concentration mixture model. The presented findings notably extend the results of [7] to a wide class of distributions of the data vectors, including Lipschitz-ally transformed Gaussian vectors. Our results notably confirm a side of *universality* of spectral clustering as suggested in [7]. More importantly, we have demonstrated, through this paper, that real data behave similar to concentrated vectors. Besides, random matrix theory allows for the theoretical understanding of machine learning methods for concentrated vectors, thereby demonstrating its relevance for machine learning.

## 6. REFERENCES


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

[2] Cosme Louart, Zhenyu Liao, and Romain Couillet, "A random matrix approach to neural networks," *The Annals of Applied Probability*, vol. 28, no. 2, pp. 1190–1248, 2018.

[3] Zhenyu Liao and Romain Couillet, "Random matrices meet machine learning: A large dimensional analysis of ls-svm," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2397–2401.

[4] Xiaoyi Mai and Romain Couillet, "A random matrix analysis and improvement of semi-supervised learning for large dimensional data," *arXiv preprint arXiv:1711.03404*, 2017.

[5] Mohamed El Amine Seddik, Mohamed Tamaazousti, and Romain Couillet, "A kernel random matrix-based approach for sparse PCA," in *International Conference on Learning Representations*, 2019.

[6] Hafiz Tiomoko Ali, Abla Kammoun, and Romain Couillet, "Random matrix-improved kernels for large dimensional spectral clustering," in *IEEE Statistical Signal Processing Workshop 2018*, 2018.

[7] Romain Couillet and Florent Benaych-Georges, "Kernel spectral clustering of large dimensional data," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1393–1454, 2016.

[8] Noureddine El Karoui et al., "The spectrum of kernel random matrices," *The Annals of Statistics*, vol. 38, no. 1, pp. 1–50, 2010.

[9] Yann LeCun, "The mnist database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*, 1998.

[10] Youssef Tamaazousti, Hervé Le Borgne, Céline Hudelot, Mohamed El Amine Seddik, and Mohamed Tamaazousti, "Learning more universal representations for transfer-learning," *arXiv:1712.09708*, 2017.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[12] Terence Tao, *Topics in random matrix theory*, vol. 132, American Mathematical Society Providence, RI, 2012.

[13] Cosme Louart and Romain Couillet, "Large sample covariance matrices of concentrated random vectors," *(in preparation)*.

[14] Alex Krizhevsky and Geoff Hinton, "Convolutional deep belief networks on cifar-10," *Unpublished manuscript*, vol. 40, no. 7, 2010.

[15] Stanley C Eisenstat and Ilse CF Ipsen, "Three absolute perturbation bounds for matrix eigenvalues imply relative bounds," *SIAM Journal on Matrix Analysis and Applications*, vol. 20, no. 1, pp. 149–158, 1998.

[16] Florent Benaych-Georges and Raj Rao Nadakuditi, "The singular values and vectors of low rank perturbations of large rectangular random matrices," *Journal of Multivariate Analysis*, vol. 111, pp. 120–135, 2012.

[17] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[18] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[19] Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael Cree, "Regularisation of neural networks by enforcing lipschitz continuity," *arXiv preprint arXiv:1804.04368*, 2018.