Kernel Random Matrices of Large Concentrated Data: the Example of GAN-Generated Images

(ENS weekly Golosino seminar)

Mohamed El Amine SEDDIK

CEA List, France
CentraleSupélec, L2S, Université ParisSaclay, France
University Paris-Saclay.

05 December 2019

Outline

Outline

In machine learning (ML),

- We are given some data

$$X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{p \times n}$$

- We aim at performing different tasks

Regression, Classification, Clustering etc.

- At the heart of these tasks, we compute **similarities**

For instance: the inner product $x_i^\intercal x_j$

Quite naturally, the Gram matrix $X^\intercal X$ appears in ML.

- **How does it behave?**

(Understating its **behavior** will let us **anticipate the performances** of a wide range of standard ML models: e.g., Ridge-Regression, LS-SVM, Spectral Clustering ...)

Outline

Behavior of the Gram Matrix for Gaussian Vectors

- Let us assume $x_i \sim \mathcal{N}(0, I_p)$

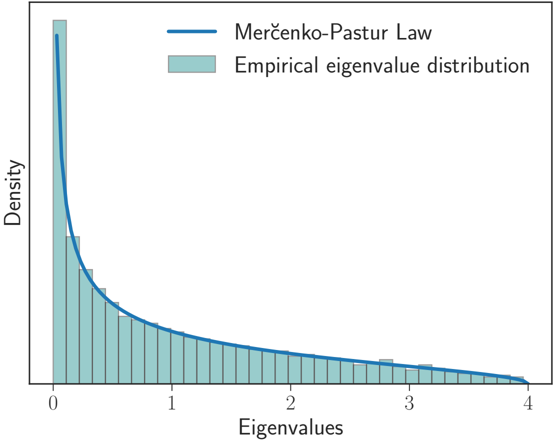


Figure: Eigenvalues distribution of $\frac{1}{p}X^\intercal X$ for $n = p = 1000$.

Definition (Empirical Spectral Density)

The empirical spectral density (e.s.d.) μ_n *of a Hermitian matrix* $A_n \in \mathbb{R}^{n \times n}$ *is given by* $\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i(A_n)}$.

Theorem (The Marčenko–Pastur Law)

Let $X \in \mathbb{R}^{p \times n}$ *with i.i.d. random entries with zero mean, and variance* 1.
When $p, n \to \infty$ *with* $n/p \to c \in (0, \infty)$, *the e.s.d.* μ_n *of* $\frac{1}{p} X^\intercal X$ *satisfies*

$$\mu_n \xrightarrow{a.s.} \mu_c$$

where μ_c *is a deterministic measure with continuous density function* f_c *on the compact support* $[\lambda^-, \lambda^+] = [(1 - \sqrt{c})^2, (1 + \sqrt{c})^2]$

$$f_c(x) = \frac{1}{2\pi c x} \sqrt{(x - \lambda^-)(\lambda^+ - x)}$$

Gaussian Mixture (Spiked Model)

- Let $\mu \in \mathbb{R}^p$ such that $\|\mu\| = \mathcal{O}(1)$
- Consider

$$X = [\underbrace{x_1, \ldots, x_{\frac{n}{2}}}_{\sim\mathcal{N}(+\mu, \mathbf{I}_p)}, \underbrace{x_{\frac{n}{2}+1}, \ldots, x_n}_{\sim\mathcal{N}(-\mu, \mathbf{I}_p)}]$$

- We can write

$$X = \mu\, y^\intercal + Z$$

where $y \in \{+1, -1\}^n$ represents the labels vector and Z has i.i.d. $\mathcal{N}(0,1)$ entries.

- We thus have

$$\frac{1}{p}X^\intercal X = \underbrace{\|\mu\|^2\, \bar{y}\bar{y}^\intercal}_{\text{Information (low-rank)}} + \underbrace{\frac{1}{p}Z^\intercal Z + *}_{\text{Noise}} \quad \text{where } \bar{\mathbf{y}} = \mathbf{y}/\sqrt{p}$$

Gaussian Mixture (Spiked Model)

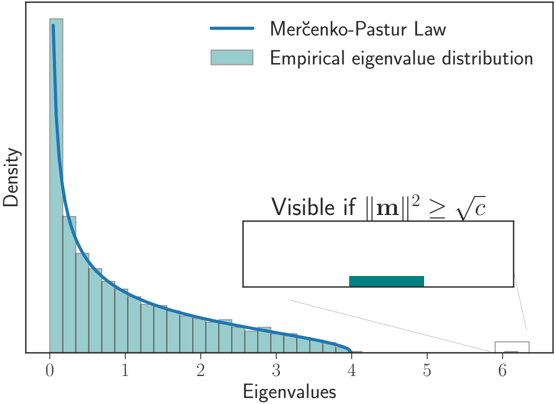


Figure: Eigenvalues distribution of $\frac{1}{p}X^\intercal X$ for $n = p = 1000$.

Gaussian Mixture (Spiked Model)

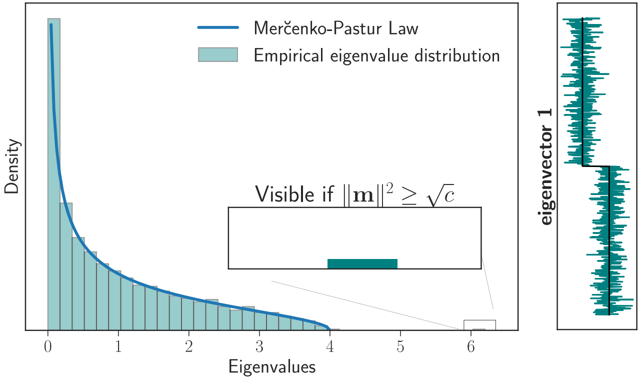


Figure: Eigenvalues distribution of $\frac{1}{p}X^{\mathsf{T}}X$ along with its dominant eigenvector for $n = p = 1000$.

Theorem ([Baik, Silverstein'06], [Paul'07])

Let

- Z *be with random i.i.d. entries with zero mean, variance* 1 *and* $\mathbb{E}|Z_{ij}|^4 < \infty$
- $X = my^\mathsf{T} + Z$

Thus, when $p, n \to$ *with* $n/p \to c$,

- *If* $\|\mu\|^2 > \sqrt{c}$

$$\lambda_\ell \left(\frac{1}{p} X^\mathsf{T} X \right) \xrightarrow{a.s.} 1 + \|\mu\|^2 + c \frac{1 + \|\mu\|^2}{\|\mu\|^2} > (1 + \sqrt{c})^2$$

- *For* $a, b \in \mathbb{R}^p$ *deterministic and* $\hat{y}$ *the eigenvector corresponding to* $\lambda_{\max}\left(\frac{1}{p}X^\mathsf{T}X\right)$,

$$a^\mathsf{T} \hat{y} \hat{y}^\mathsf{T} b - \frac{1 - c\|\mu\|^{-4}}{1 + c\|\mu\|^{-2}} a^\mathsf{T} \hat{y} \hat{y}^\mathsf{T} b \cdot 1_{\|\mu\|^2 > \sqrt{c}} \xrightarrow{a.s.} 0$$

In particular,

$$|\hat{y}^\mathsf{T} y|^2 \xrightarrow{a.s.} \frac{1 - c\|\mu\|^{-4}}{1 + c\|\mu\|^{-2}} \cdot 1_{\|\mu\|^2 > \sqrt{c}}.$$

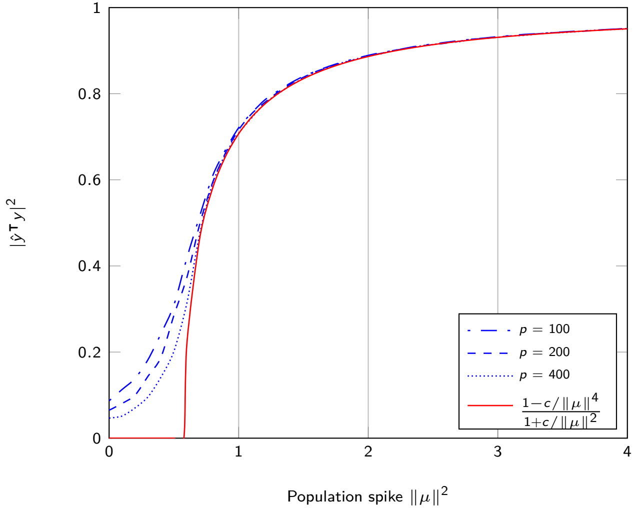


Figure: Simulated $|\hat{y}^\mathsf{T} y|^2$ and limit values, for $p/n = 1/3$, and varying $\|\mu\|^2$.

Outline

Outline

- **Observation**: RMT seems to predict ML performances in high-dimension based on Gaussian assumptions on the data.
- **BUT** Real Data are **unlikely close** to Gaussian vectors!
- Gaussian vectors fall within a larger, more useful, class of random vectors!

Definition

Given a normed space $(E, \|\cdot\|_E)$ et $q \in \mathbb{R}$, a random vector $\mathbf{z} \in E$ is q-exponentially **concentrated** if for any 1-**Lipschitz**[1] function $\mathcal{F} : \mathbb{R}^p \to \mathbb{R}$, there exists $C, c > 0$ such that

$$\mathbb{P}\left\{|\mathcal{F}(\mathbf{z}) - \mathbb{E}\mathcal{F}(\mathbf{z})| > t\right\} \leq Ce^{-c\,t^q} \xrightarrow{\text{denoted}} \boxed{\mathbf{z} \in \mathcal{O}(e^{-\cdot^q})}$$

(P1) $X \sim \mathcal{N}(0, I_p)$ is 2-exponentially **concentrated**.

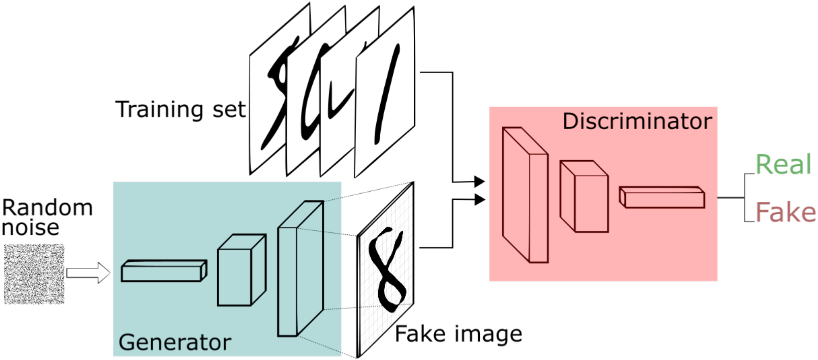
(P2) If $X \in \mathcal{O}(e^{-\cdot^q})$ and $\mathcal{G}$ is a $\|\mathcal{G}\|_{lip}$-**Lipschitz** transformation, then

$$\mathcal{G}(X) \in \mathcal{O}\left(e^{-(\cdot/\|\mathcal{G}\|_{lip})^q}\right).$$

"Concentrated vectors are stable through Lipschitz maps."

[1] **Reminder:** $\mathcal{F} : E \to F$ is $\|\mathcal{F}\|_{lip}$-Lipschitz if $\forall (x, y) \in E^2 : \|\mathcal{F}(x) - \mathcal{F}(y)\|_F \leq \|\mathcal{F}\|_{lip}\,\|x - y\|_E$.

Outline



$$\min_{\mathcal{G}} \max_{\mathcal{D}} \mathbb{E}_{x \sim p(x)}[\log \mathcal{D}(x)] + \mathbb{E}_{z \sim p(z)}[\log(1 - \mathcal{D}(\mathcal{G}(z)))]$$

Once the Generator is trained, we generate data as

$$\boxed{\text{Generated Image} = \mathcal{G}(\text{Gaussian})}$$

GAN Data: An Example of Concentrated Vectors

Figure: Images artificially generated using the BigGAN model [Brock *et al*, ICLR'19].

$$\text{GAN Data} = \mathcal{F}_1 \circ \mathcal{F}_2 \circ \cdots \circ \mathcal{F}_N(\text{Gaussian})$$

where the $\mathcal{F}_i$'s correspond to Fully Connected layers, Convolutional layers, Pooling and activation functions, residual connections or Batch Normalisation.

$\Rightarrow$ The $\mathcal{F}_i$'s are essentially *Lipschitz* operations.

- **Fully Connected Layers and Convolutional Layers** are affine operations:

$$\mathcal{F}_i(x) = W_i x + b_i,$$

  and $\|\mathcal{F}_i\|_{lip} = \sup_{u \neq 0} \frac{\|W_i u\|_p}{\|u\|_p}$, for any p-norm.

- **Pooling Layers and Activation Functions:** Are 1-Lipschitz operations with respect to any p-norm (e.g., ReLU and Max-pooling).

- **Residual Connections:** $\mathcal{F}_i(x) = x + \mathcal{F}_i^{(1)} \circ \cdots \circ \mathcal{F}_i^{(\ell)}(x)$
  where the $\mathcal{F}_i^{(j)}$'s are Lipschitz operations, thus $\mathcal{F}_i$ is a Lipschitz operation with Lipschitz constant bounded by $1 + \prod_{j=1}^{\ell} \|\mathcal{F}_i^{(j)}\|_{lip}$.

- ...

Consider data distributed in k classes $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ as

$$X = [\underbrace{x_1, \ldots, x_{n_1}}_{\in \mathcal{O}(e^{-\cdot^{q_1}})}, \underbrace{x_{n_1+1}, \ldots, x_{n_2}}_{\in \mathcal{O}(e^{-\cdot^{q_2}})}, \ldots, \underbrace{x_{n-n_k+1}, \ldots, x_n}_{\in \mathcal{O}(e^{-\cdot^{q_k}})}] \in \mathbb{R}^{p \times n}$$

Denote

$$\mu_\ell = \mathbb{E}_{x_i \in \mathcal{C}_\ell}[x_i], \quad C_\ell = \mathbb{E}_{x_i \in \mathcal{C}_\ell}[x_i x_i^\intercal]$$

Assumption (Growth rate)

As $p \to \infty$,

1. $p/n \to c \in (0, \infty)$.
2. *The number of classes k is bounded.*
3. *For any $\ell \in [k]$, $\|\mu_\ell\| = \mathcal{O}(\sqrt{p})$.*

Outline

Let

$$G = \frac{1}{p} X^\intercal X = \frac{1}{p} J M^\intercal M J^\intercal + \frac{1}{p} Z^\intercal Z + * + o_p(1)$$

Denote by L the e.s.d. of G and U the matrix containing the top dominant eigenvectors of G. Then

$$L = \frac{1}{n} \sum_{i}^{n} \delta_{\lambda_i},\ m_L(z) = \int_\lambda \frac{dL(\lambda)}{\lambda - z} = \frac{1}{n} \mathrm{tr}\left(Q(-z)\right)$$

$$UU^\intercal = \frac{1}{2\pi i} \oint_\gamma Q(-z) dz$$

$$\boxed{\Rightarrow \text{Analyse the behavior of the resolvent } Q(z) = (G + zI_n)^{-1}.}$$

Theorem

Under the assumptions above, we have $Q(z) \in \mathcal{O}(e^{-(\sqrt{p}\,\cdot)^q})$ in $(\mathbb{R}^{n\times n}, \|\cdot\|)$. Furthermore,

$$\left\|\mathbb{E}[Q(z)] - \tilde{Q}(z)\right\| = \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right) \text{ where } \tilde{Q}(z) = \frac{1}{z}\Lambda(z) + \frac{1}{p\,z}J\Omega(z)J^\intercal$$

with $\Lambda(z) = diag\left\{\frac{1_{n_\ell}}{1+\delta_\ell(z)}\right\}_{\ell=1}^k$ and $\Omega(z) = diag\{\mu_\ell^\intercal \tilde{R}(z)\mu_\ell\}_{\ell=1}^k$

$$\tilde{R}(z) = \left(\frac{1}{k}\sum_{\ell=1}^k \frac{C_\ell}{1+\delta_\ell(z)} + zI_p\right)^{-1}$$

with $\delta(z) = [\delta_1(z), \ldots, \delta_k(z)]$ is the unique fixed point of the system of equations

$$\delta_\ell(z) = tr\left(C_\ell\left(\frac{1}{k}\sum_{j=1}^k \frac{C_j}{1+\delta_j(z)} + zI_p\right)^{-1}\right) \text{ for each } \ell \in [k].$$

Theorem

Under the assumptions above, we have $Q(z) \in \mathcal{O}(e^{-(\sqrt{p}\,\cdot)^q})$ in $(\mathbb{R}^{n\times n}, \|\cdot\|)$. Furthermore,

$$\left\|\mathbb{E}[Q(z)] - \tilde{Q}(z)\right\| = \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right) \quad \textit{where} \;\; \tilde{R}(z) = \frac{1}{z}\Lambda(z) + \frac{1}{p\,z} J\Omega(z)J^{\intercal}$$

with $\Lambda(z) = diag\left\{\frac{1_{n_\ell}}{1+\delta_\ell(z)}\right\}_{\ell=1}^{k}$ *and* $\Omega(z) = diag\{\mu_\ell{}^{\intercal}\tilde{R}(z)\mu_\ell\}_{\ell=1}^{k}$

$$\tilde{R}(z) = \left(\frac{1}{k}\sum_{\ell=1}^{k}\frac{C_\ell}{1+\delta_\ell(z)} + zI_p\right)^{-1}$$

with $\delta(z) = [\delta_1(z), \ldots, \delta_k(z)]$ *is the unique fixed point of the system of equations*

$$\delta_\ell(z) = tr\left(C_\ell\left(\frac{1}{k}\sum_{j=1}^{k}\frac{C_j}{1+\delta_j(z)} + zI_p\right)^{-1}\right) \quad \textit{for each} \;\; \ell \in [k].$$

Key Observation: Only first and second order statistics matter!

Outline

Problem Statement

- Given data $x_1, \ldots, x_n \in \mathbb{R}^p$
- Objective: "cluster" in k **similarity** classes.
- Based on a kernel matrix

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

Intuition (from small dimensions)

$$\mathbf{K} = \begin{array}{cc} \begin{pmatrix} \gg 1 & \ll 1 & \ll 1 \\ \ll 1 & \gg 1 & \ll 1 \\ \ll 1 & \ll 1 & \gg 1 \end{pmatrix} & \begin{array}{c} \mathcal{C}_1 \\ \mathcal{C}_2 \\ \mathcal{C}_3 \end{array} \end{array}$$

- $\mathbf{K}$ mainly low rank with class information in eigenvectors.

Small Dimension vs High Dimension!

$$\mathbf{K} = \left(\begin{array}{ccc} \gg 1 & \ll 1 & \ll 1 \\ \ll 1 & \gg 1 & \ll 1 \\ \ll 1 & \ll 1 & \gg 1 \end{array}\right) \begin{array}{c} \mathcal{C}_1 \\ \mathcal{C}_2 \\ \mathcal{C}_3 \end{array}$$

$$\mathbf{K} = \left(\begin{array}{ccc} \gg 1 & \gg 1 & \gg 1 \\ \gg 1 & \gg 1 & \gg 1 \\ \gg 1 & \gg 1 & \gg 1 \end{array}\right) \begin{array}{c} \mathcal{C}_1 \\ \mathcal{C}_2 \\ \mathcal{C}_3 \end{array}$$

- **Key Observation:** *The between and within class vectors are "equidistant" in high-dimension.*

$$\boxed{\max_{1\leq i\neq j\leq n}\left\{\left|\frac{1}{p}\|x_i-x_j\|^2-\tau\right|\right\}=\mathcal{O}\left(\frac{\log(\frac{p}{\sqrt{\delta}})^{1/q}}{\sqrt{p}}\right)\to 0}$$

where $\tau=\frac{2}{p}\operatorname{tr}C$, and $C=\sum_{\ell=1}^{k}\frac{n_\ell}{n}C_\ell$.

- Taylor Expanding $\mathbf{K}$ entry-wise leads to

$$\mathbf{K}\propto\underbrace{\mathbf{J}\mathbf{A}\mathbf{J}^\intercal}_{\textit{Information}}+\underbrace{f'(\tau)\mathbf{Z}^\intercal\mathbf{Z}+*}_{\textit{Noise}}$$

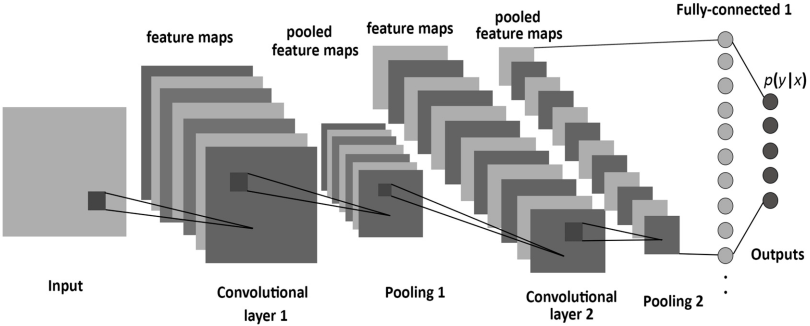
where $\mathbf{A}\propto f'(\tau)\mathbf{M}^\intercal\mathbf{M}+f''(\tau)[\mathbf{t}\mathbf{t}^\intercal+\mathbf{T}]$,and

$$\mathbf{J}=[\mathbf{j}_1,\ldots,\mathbf{j}_k],\ \mathbf{M}=[\bar{\mathbf{m}}_1,\ldots,\bar{\mathbf{m}}_k]\ \ \mathbf{t}=\left\{\frac{\operatorname{tr}\bar{\mathbf{C}}_\ell}{\sqrt{p}}\right\}_{\ell=1}^{k},\ \mathbf{T}=\left\{\frac{\operatorname{tr}\bar{\mathbf{C}}_a\bar{\mathbf{C}}_b}{p}\right\}_{a,b=1}^{k}$$

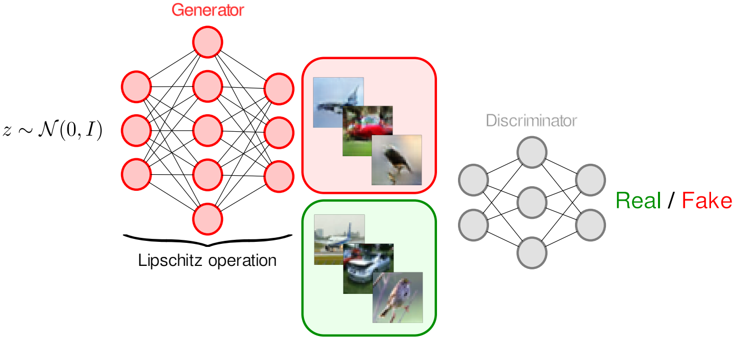
Result: Only first and second order statistics matter!

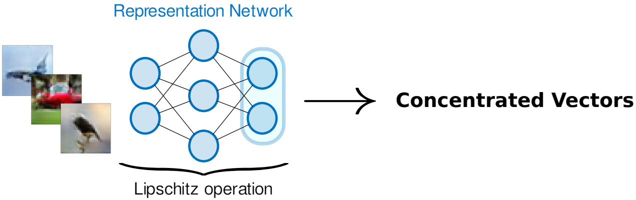
Outline

Application to CNN Representations of GAN Images



- CNN representations correspond to the **one before last** layer.
- Popular architectures considered in practice are: **Resnet, VGG, Densenet**.

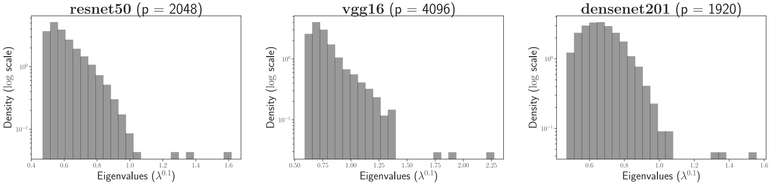

Generator
$z \sim \mathcal{N}(0, I)$
Lipschitz operation
Discriminator
Real / Fake


Representation Network
Lipschitz operation
Concentrated Vectors

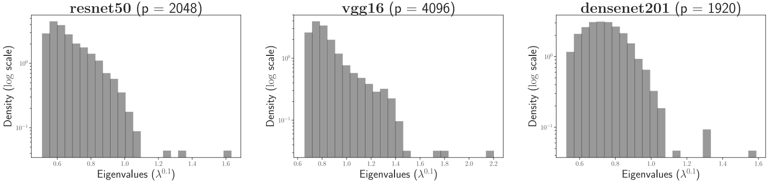
Application to CNN Representations of GAN Images

Application to CNN Representations of GAN Images

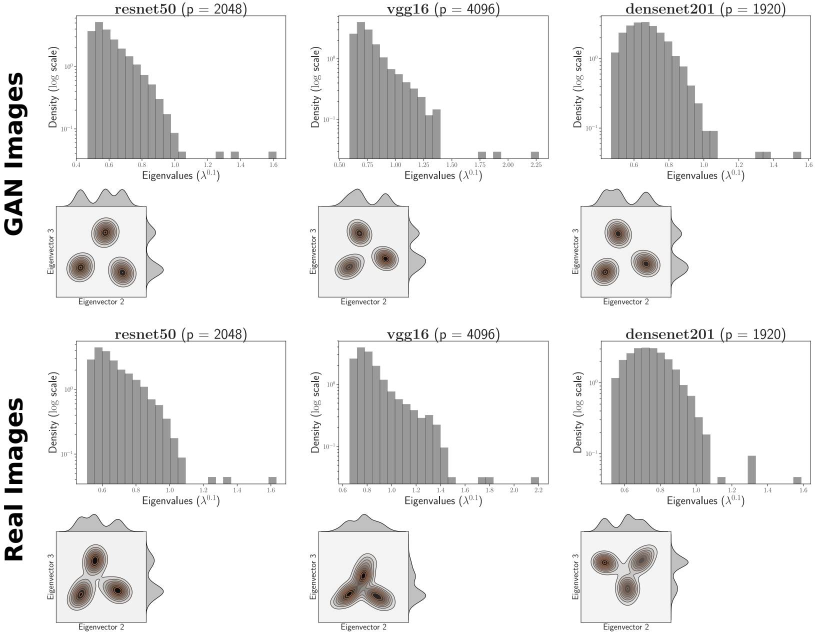
GAN Images



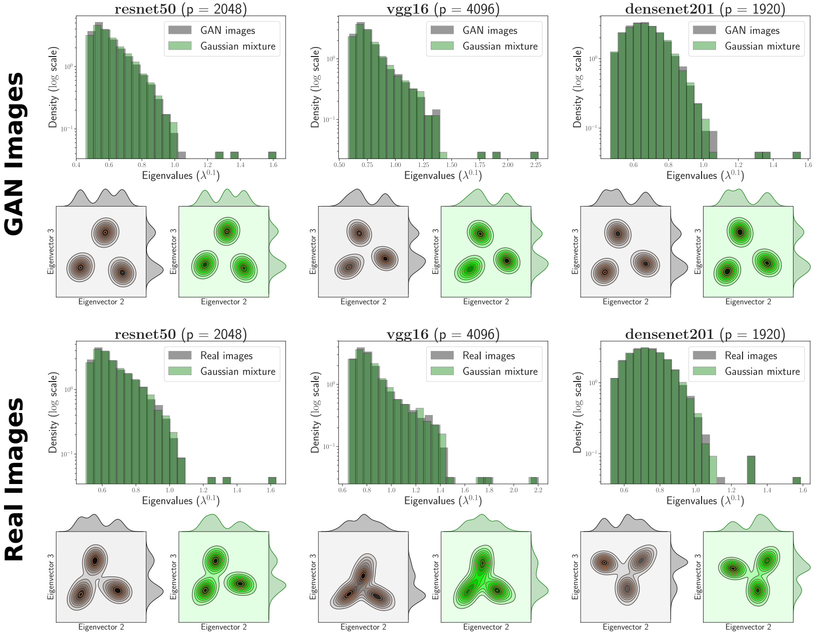
Real Images



Application to CNN Representations of GAN Images



Application to CNN Representations of GAN Images



Perspectives

- Extensions to other ML methods (SVM, SSL ... etc).
- Considering ML algorithms with implicit solutions (last layer of a neural network).
- Definition of a criterion for choosing the best representation in a Transfer-Learning framework.
- Use of the concentration of measure framework for improving GAN generation and entropy.

Thanks for your attention!

Web-page: http://melaseddik.github.io/