Random Matrix Theory for AI: From Theory to Practice

Ph.D. defense

Mohamed El Amine Seddik
supervised by **Romain Couillet** & **Mohamed Tamaazousti**

https://melaseddik.github.io/

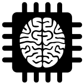
CEA List, CentraleSupélec, University of Paris-Saclay, France

November 3, 2020

Abstract

High-dimensional Data

$x_1, \ldots, x_n \in \mathbb{R}^p$



Machine Learning

Context:

- Study of standard ML classifiers on **real** **high-dimensional** data.

Motivation:

- RMT predicts performances under **Gaussian** data model.
- **BUT Real data** are **unlikely close** to **Gaussian** vectors.

In this thesis, we highlighted:

- **GAN data (≈ Real data)** are **Concentrated** vectors.
- **Universality result:**

> *Only* **first** and **second** order statistics of **Concentrated** data describe behavior of studied classifiers.

Outline

Outline

Large Sample Covariance Matrices (MP'67)

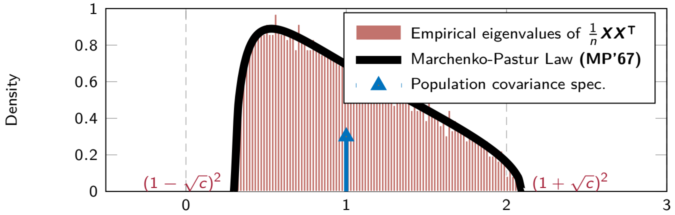
- Let $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{p \times n}$ such that $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$.
- Maximum likelihood suggests **sample covariance** as estimator for **population covariance** (here $\boldsymbol{C} = \boldsymbol{I}_p$).

$$\hat{\boldsymbol{C}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^{\intercal} = \frac{1}{n} \boldsymbol{X}\boldsymbol{X}^{\intercal} \xrightarrow{\text{a.s.}} \boldsymbol{I}_p$$

consistent when $n \to \infty$ with p fixed.

- When $p \sim n$, inconsistency occurs:

$$\|\hat{\boldsymbol{C}} - \boldsymbol{I}_p\| \not\to 0 \quad \text{as} \quad n, p \to \infty, \ \frac{p}{n} \to c \in (0, \infty)$$



Example of drawback: $\frac{1}{p}\|\boldsymbol{C}\|_F^2 = \frac{1}{p}\operatorname{tr}(\boldsymbol{C}^2) \approx \frac{1}{p}\operatorname{tr}(\hat{\boldsymbol{C}}^2) - c\left(\frac{1}{p}\operatorname{tr}(\hat{\boldsymbol{C}})\right)^2$.

Large Kernel Matrices (ElK'10, CBG'16)

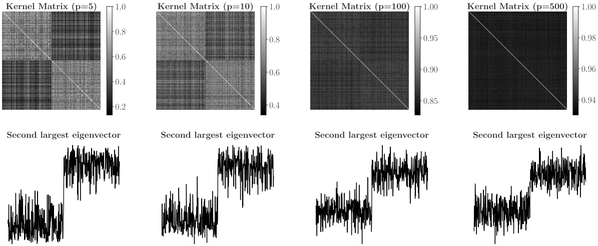
- Let $\boldsymbol{x}_i = \left\{ \begin{array}{c} +\boldsymbol{\mu} \\ \text{or} \\ -\boldsymbol{\mu} \end{array} \right\} + \boldsymbol{z}_i$ with $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$.
- Separability **possible only** if $\|\boldsymbol{\mu}\| \geq \mathcal{O}(1)$ by Neyman-Pearson test.
- Implies (in worst case) **non-trivial** growth setting

$$\max_{1 \leq i \neq j \leq n} \left\{ \frac{1}{p} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 - 2 \right\} \xrightarrow{\text{a.s.}} 0 \quad \text{as} \quad p \to \infty$$

irrespective of classes ($\mathcal{C}_1$ or $\mathcal{C}_2$) of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

- Taylor expanding $K_{ij} \equiv f\left(\frac{1}{p}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right)$ yields (for $\boldsymbol{j} \equiv [+\boldsymbol{1}_{\frac{n}{2}}, -\boldsymbol{1}_{\frac{n}{2}}]$)

$$\boldsymbol{K} = f(2)\boldsymbol{1}_n\boldsymbol{1}_n^\intercal + f'(2)(\boldsymbol{Z}^\intercal\boldsymbol{Z}/p + \varphi(\boldsymbol{\mu})\boldsymbol{j}\boldsymbol{j}^\intercal/p) + * \quad \text{as} \quad \frac{p}{n} \to c \in (0, \infty)$$



RMT Meets Machine Learning

RMT predicts **performances** of various ML methods:

- Kernel Spectral Clustering **(Couillet+'16)**.
- Least Squares Support Vectors Machines **(Liao+'17)**.
- Semi-supervised Learning **(Mai+'17)**.
- Random Shallow Neural Networks **(Pennington+'17, Louart+'18)**.
- Random Feature Maps **(Liao+'18)**.
- Learning Dynamics of Shallow Nets **(Liao+'18)**.
- Loss Surface Geometry of Deep nets **(Choromanska+'15, Pennington+'17)**.
- Learning with Dropout **(Seddik+'20)**.
- Analysis of Logistic Regression **(ElKaroui+'13, Mai+'19)**.
- Multi-task and Transfer Learning **(Tiomoko+'20)**.

Mostly under Gaussian assumptions (for $\boldsymbol{x}_i \in \mathcal{C}_\ell$):

$$\boldsymbol{x}_i = \boldsymbol{\mu}_\ell + \boldsymbol{\Sigma}_\ell^{\frac{1}{2}} \boldsymbol{z}_i \quad \text{with} \quad \boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$$

Real data

$\neq$

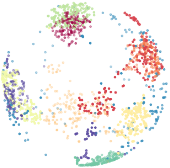
Gaussian data

$$= \hat{\mu}_2 + \hat{\Sigma}_2^{\frac{1}{2}} z_i \quad \text{with} \quad z_i \sim \mathcal{N}(0, I_p)$$

Outline

From GMMs to Concentration through GANs

$$z \sim \mathcal{N}(0, I_p) \to \mathcal{G}(z) \to$$



$x_1, \ldots, x_n \in \mathbb{R}^p$

Contribution 1

GAN-data: Example of Concentrated Vectors

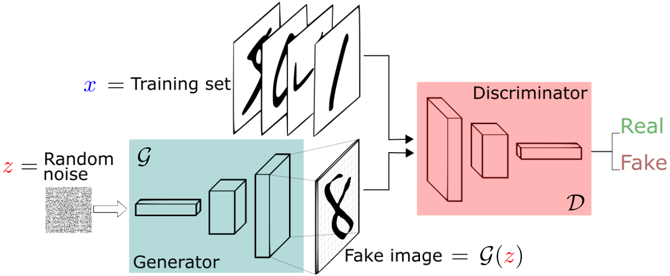
*MEA. Seddik, C. Louart, M. Tamaazousti, R. Couillet, "**Random Matrix Theory Proves that Deep Learning Representations of GAN-data Behave as Gaussian Mixtures**", ICML'2020.*

From GMMs to Concentration through GANs

- Following R. Feynman's quote:

> "***What I cannot create, I do not understand***"

- Generative models provide examples of **realistic data**.



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

Generated images $= \mathcal{G}($**Gaussian**$)$

From GMMs to Concentration through GANs

Figure: Images artificially generated with BigGAN **(BDS'19)**.

$$\textbf{Real Data} \approx \textbf{GAN Data} = \underbrace{\Phi_L \circ \Phi_{L-1} \circ \cdots \circ \Phi_1}_{\mathcal{G}}(\textbf{Gaussian})$$

where Φ_i's correspond to standard NN operations.

$\Rightarrow$ The Φ_i's are ***Lipschitz*** maps.

From GMMs to Concentration through GANs

Definition (Concentrated Vectors)

Given a normed space $(\mathcal{X}, \|\cdot\|)$ and $q > 0$, a random vector $\boldsymbol{x} \in \mathcal{X}$ is q-exponentially **concentrated** if for any 1-**Lipschitz** function $\varphi : \mathcal{X} \to \mathbb{R}$, there exist $C, \sigma > 0$ such that

$$\forall t > 0,\ \mathbb{P}\left\{|\varphi(\boldsymbol{x}) - \mathbb{E}\varphi(\boldsymbol{x})| \geq t\right\} \leq Ce^{-(t/\sigma)^q} \xrightarrow{\text{denoted}} \boxed{\boldsymbol{x} \propto \mathcal{E}_q(\sigma)}$$

If σ independent of $\dim(\mathcal{X})$, we denote $\boxed{\boldsymbol{x} \propto \mathcal{E}_q}$.

Concentrated vectors enjoy:

(P1) If $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$ then $\boldsymbol{z} \propto \mathcal{E}_2$

"Gaussian vectors are concentrated vectors"

(P2) If $\boldsymbol{z} \propto \mathcal{E}_q$ and $\mathcal{G}$ is a $\lambda_{\mathcal{G}}$-**Lipschitz** map, then $\mathcal{G}(\boldsymbol{z}) \propto \mathcal{E}_q(\lambda_{\mathcal{G}})$

"Concentrated vectors are stable through Lipschitz maps"

$\Rightarrow$ **GAN data** are **concentrated** vectors by design.

Remark: Still, we need to control $\lambda_{\mathcal{G}}$.

Control of $\lambda_{\mathcal{G}}$ with Spectral Normalization (SN)

- SN stabilizes learning of GANs **(BD+'19)**.
- SN makes neural nets robust against adversarial examples **(SZ+'13, AS+'17)**.

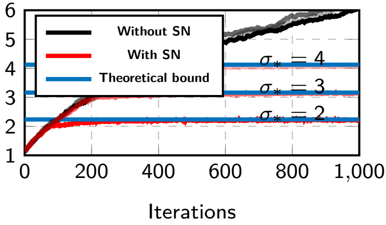
- Let $\sigma_* > 0$ and $\mathcal{G}$ a N-layers NN
- d_{i-1}: input dim, d_i: output dim of layer i
- Assimilate SGD to random walk **(AS'18)**:

$$\boldsymbol{W} \leftarrow \boldsymbol{W} - \eta \boldsymbol{E}, \text{ with } \boldsymbol{E}_{i,j} \sim \mathcal{N}(0,1)$$

$$\boldsymbol{W} \leftarrow \boldsymbol{W} - \max(0, \sigma_1(\boldsymbol{W}) - \sigma_*)\, \boldsymbol{u}_1(\boldsymbol{W}) \boldsymbol{v}_1(\boldsymbol{W})^\intercal \quad \textbf{(with SN)}$$

$\lambda_{\mathcal{G}}$ bounded **(SLTC'20)**, for $\varepsilon > 0$

$$\lambda_{\mathcal{G}} \leq \prod_{i=1}^{N} \left(\varepsilon + \sqrt{\sigma_*^2 + \eta^2 d_i d_{i-1}} \right)$$



Random Matrix Theory for AI: From Theory to Practice

MEA. Seddik

Some ML methods under Concentration

$x_1, \ldots, x_n \in \mathbb{R}^p$

$\longrightarrow$

$$G_{ij} = \frac{1}{p} x_i^\intercal x_j$$

Contribution 2

Linear Classifiers: Behavior of Gram Matrices

MEA. Seddik, *C. Louart, M. Tamaazousti, R. Couillet, "**Random Matrix Theory Proves that Deep Learning Representations of GAN-data Behave as Gaussian Mixtures**", ICML'2020.*

Model & Assumptions

(A1) Data matrix (distributed in k classes $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$):

$$\boldsymbol{X} = \left[\underbrace{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_1}}_{\propto \mathcal{E}_{q_1}}, \underbrace{\boldsymbol{x}_{n_1+1}, \ldots, \boldsymbol{x}_{n_2}}_{\propto \mathcal{E}_{q_2}}, \ldots, \underbrace{\boldsymbol{x}_{n-n_k+1}, \ldots, \boldsymbol{x}_n}_{\propto \mathcal{E}_{q_k}} \right] \in \mathbb{R}^{p \times n}$$

Model statistics: $\boldsymbol{\mu}_\ell = \mathbb{E}_{\boldsymbol{x}_i \in \mathcal{C}_\ell}[\boldsymbol{x}_i], \quad \boldsymbol{\Sigma}_\ell = \mathbb{E}_{\boldsymbol{x}_i \in \mathcal{C}_\ell}[\boldsymbol{x}_i \boldsymbol{x}_i^\intercal] - \boldsymbol{\mu}_\ell \boldsymbol{\mu}_\ell^\intercal$

(A2) Growth rate assumptions: As $p \to \infty$,

1. $p/n \to c \in (0, \infty)$.
2. k fixed.
3. $\|\boldsymbol{\mu}_\ell\| = \mathcal{O}(\sqrt{p})$.

Gram matrix and its resolvent:

$$\boldsymbol{G} = \frac{1}{p} \boldsymbol{X}^\intercal \boldsymbol{X}, \quad \boldsymbol{Q}(z) = (\boldsymbol{G} + z \boldsymbol{I}_n)^{-1}$$

$$m(z) = \frac{1}{n} \mathrm{tr}\left(\boldsymbol{Q}(-z)\right), \quad \boldsymbol{U}\boldsymbol{U}^\intercal = \frac{-1}{2\pi i} \oint_\gamma \boldsymbol{Q}(-z) dz$$

Notion of *Deterministic Equivalent*

Definition (Deterministic Equivalent **(Hachem+'07)**)

$$\boldsymbol{Q} \leftrightarrow \bar{\boldsymbol{Q}}$$

if for all $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ of bounded norms:

$$\frac{1}{n}\operatorname{tr} \boldsymbol{A}(\boldsymbol{Q} - \bar{\boldsymbol{Q}}) \xrightarrow{\text{a.s.}} 0, \quad \boldsymbol{a}^\intercal(\boldsymbol{Q} - \bar{\boldsymbol{Q}})\boldsymbol{b} \xrightarrow{\text{a.s.}} 0$$

Examples (Sample covariance matrix **(Louart+'18)**)

Let $k = 1$ and $\boldsymbol{C} = \boldsymbol{\Sigma}_1 + \boldsymbol{\mu}_1\boldsymbol{\mu}_1^\intercal$

$$\boldsymbol{R}(z) \equiv \left(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\intercal + z\boldsymbol{I}_p\right)^{-1} \leftrightarrow \bar{\boldsymbol{R}}(z) \equiv \left(\frac{\boldsymbol{C}}{1+\delta} + z\boldsymbol{I}_p\right)^{-1} \qquad \delta = \frac{1}{n}\operatorname{tr}\left(\boldsymbol{C}\bar{\boldsymbol{R}}(z)\right)$$

For $\bar{\boldsymbol{R}}(z) = (\boldsymbol{F} + z\boldsymbol{I}_p)^{-1}$:

$$\tilde{\boldsymbol{R}} - \mathbb{E}\boldsymbol{R} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\boldsymbol{R}_{-i}\left(\frac{\boldsymbol{x}_i\boldsymbol{x}_i^\intercal}{1 + \frac{1}{n}\boldsymbol{x}_i^\intercal\boldsymbol{R}_{-i}\boldsymbol{x}_i} - \boldsymbol{F}\right)\bar{\boldsymbol{R}}\right] + *$$

Remark: $\delta = 0$ in the classical regime: $n \to \infty$ with p fixed.

Main Result: Universality of Linear Classifiers

Theorem (Resolvent of the Gram Matrix (SLTC'20))

Under Assumptions (A1-2), we have $\boldsymbol{Q}(z) \propto \mathcal{E}_q(p^{-\frac{1}{2}})$. *Furthermore,*

$$\boldsymbol{Q}(z) \leftrightarrow \bar{\boldsymbol{Q}}(z) \equiv \frac{1}{z}\boldsymbol{\Lambda}(z) + \frac{1}{p\,z}\boldsymbol{J}\Omega(z)\boldsymbol{J}^\intercal$$

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

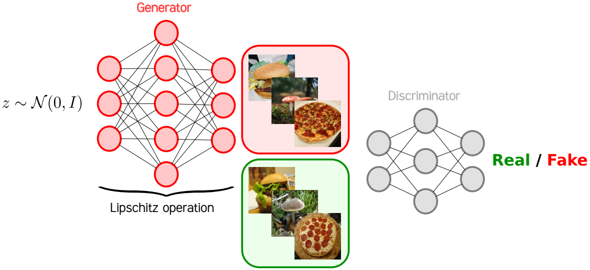
$$\bar{\boldsymbol{R}}(z) = \left(\frac{1}{k}\sum_{\ell=1}^k \frac{\boldsymbol{\Sigma}_\ell + \boldsymbol{\mu}_\ell\boldsymbol{\mu}_\ell^\intercal}{1+\delta_\ell(z)} + z\boldsymbol{I}_p\right)^{-1}$$

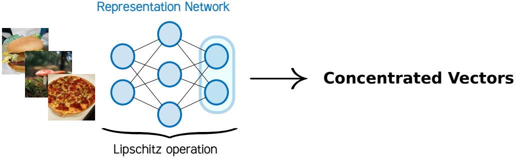
with $\delta(z) = [\delta_1(z), \ldots, \delta_k(z)]$ *unique solution to:*

$$\delta_\ell(z) = tr\left((\boldsymbol{\Sigma}_\ell + \boldsymbol{\mu}_\ell\boldsymbol{\mu}_\ell^\intercal)\left(\frac{1}{k}\sum_{j=1}^k \frac{\boldsymbol{\Sigma}_j + \boldsymbol{\mu}_j\boldsymbol{\mu}_j^\intercal}{1+\delta_j(z)} + z\boldsymbol{I}_p\right)^{-1}\right) \quad \text{for each } \ell \in [k]$$

Key Observation: Only **first** and **second** order statistics matter!

Application to CNN Representations of GAN Images



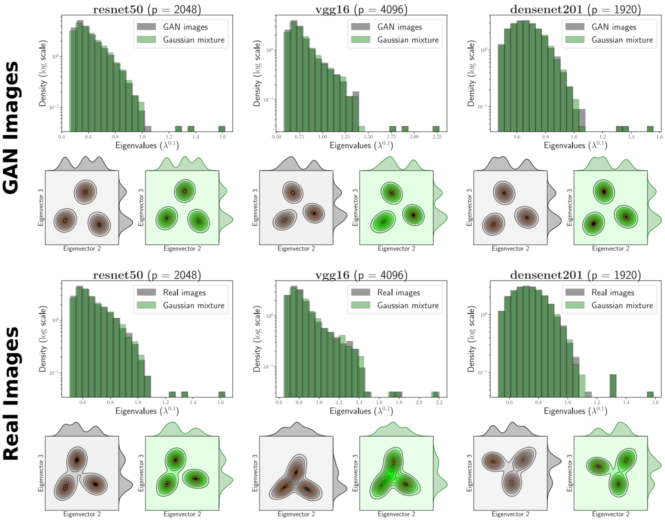


- CNN representations $\rightarrow$ **penultimate** layer.
- Popular architectures: **Resnet, VGG, Densenet**.

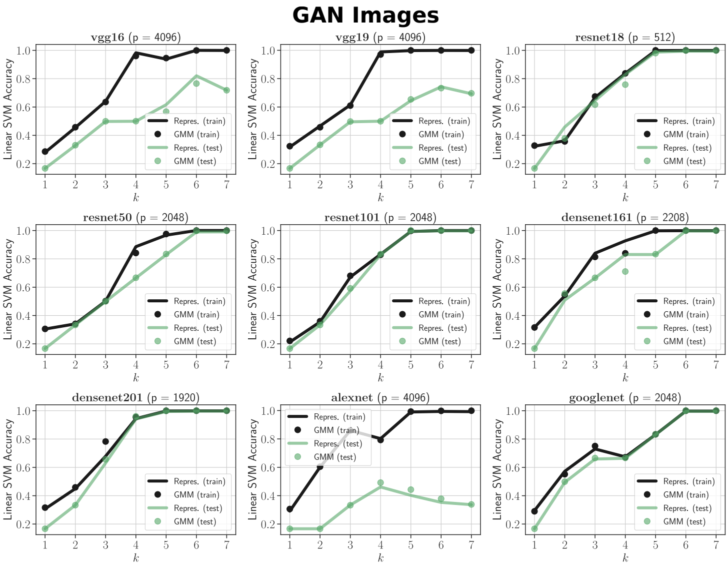
Application to CNN Representations of GAN Images

Figure: $k = 3$ classes, $n = 3000$ images.

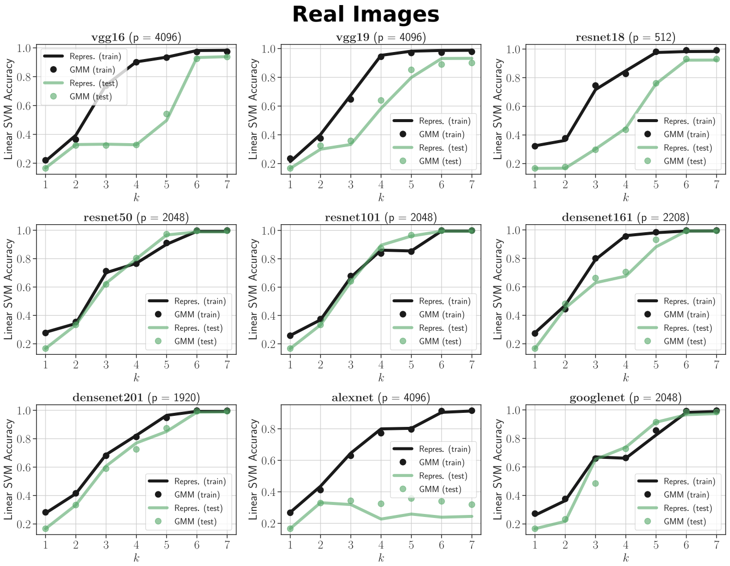
Application to CNN Representations of GAN Images

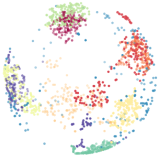


Performance of a linear SVM classifier (GAN data)



Performance of a linear SVM classifier (Real data)





$x_1, \ldots, x_n \in \mathbb{R}^p$

$$\longrightarrow \quad \boxed{\kappa(x_i, x_j) = f\left(\frac{1}{p}\|x_i - x_j\|^2\right)}$$

Contribution 3

Kernel Methods: Behavior of Kernel Matrices

MEA. Seddik, *M. Tamaazousti, R. Couillet, "**Kernel Random Matrices of Large Concentrated Data: The Example of GAN-generated Images**", ICASSP'2019.*

Model & Assumptions

(A1) Data matrix (distributed in k classes $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$):

$$\boldsymbol{X} = \left[\underbrace{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_1}}_{\propto \mathcal{E}_{q_1}}, \underbrace{\boldsymbol{x}_{n_1+1}, \ldots, \boldsymbol{x}_{n_2}}_{\propto \mathcal{E}_{q_2}}, \ldots, \underbrace{\boldsymbol{x}_{n-n_k+1}, \ldots, \boldsymbol{x}_n}_{\propto \mathcal{E}_{q_k}} \right] \in \mathbb{R}^{p \times n}$$

Model statistics: $\boldsymbol{\mu}_\ell = \mathbb{E}_{\boldsymbol{x}_i \in \mathcal{C}_\ell}[\boldsymbol{x}_i], \quad \boldsymbol{\Sigma}_\ell = \mathbb{E}_{\boldsymbol{x}_i \in \mathcal{C}_\ell}[\boldsymbol{x}_i \boldsymbol{x}_i^\intercal] - \boldsymbol{\mu}_\ell \boldsymbol{\mu}_\ell^\intercal$

$$\boldsymbol{\mu} = \sum_{\ell=1}^k \frac{n_\ell}{n} \boldsymbol{\mu}_\ell, \quad \bar{\boldsymbol{\mu}}_\ell = \boldsymbol{\mu} - \boldsymbol{\mu}_\ell, \quad \boldsymbol{\Sigma} = \sum_{\ell=1}^k \frac{n_\ell}{n} \boldsymbol{\Sigma}_\ell, \quad \bar{\boldsymbol{\Sigma}}_\ell = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_\ell$$

(A2) Growth rate assumptions: As $p \to \infty$,

- **(Data)** $p/n \to c \in (0, \infty)$, $n_\ell / n \to c_\ell \in (0,1)$, k fixed.
- **(Means)** $\|\bar{\boldsymbol{\mu}}_\ell\| = \mathcal{O}(1)$.
- **(Covariances)** $\|\bar{\boldsymbol{\Sigma}}_\ell\| = \mathcal{O}(1)$, $\operatorname{tr} \bar{\boldsymbol{\Sigma}}_\ell = \mathcal{O}(\sqrt{p})$.

(A3) Kernel function: Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ 3-times differentiable at $\tau = \frac{2}{p} \operatorname{tr} \boldsymbol{\Sigma}$.

Kernel matrix:

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

Main Result: Universality of Kernel Matrices

We still have:

Denote $\tau \equiv \frac{2}{p}\operatorname{tr}\boldsymbol{\Sigma}$. Under **(A1-2)**, with probability $1-\delta$

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

irrespective of classes of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

$\boldsymbol{M}=[\bar{\boldsymbol{\mu}}_1,\ldots,\bar{\boldsymbol{\mu}}_k]\in\mathbb{R}^{p\times k}$, $\boldsymbol{Z}=\boldsymbol{X}-\boldsymbol{M}\boldsymbol{J}^\intercal\in\mathbb{R}^{p\times n}$ and $\boldsymbol{J}=[\boldsymbol{j}_1,\ldots,\boldsymbol{j}_k]\in\mathbb{R}^{n\times k}$

Theorem (Random Matrix Equivalent for $\boldsymbol{K}$ (STC'19))

*Under **(A1-3)** Taylor expanding $\boldsymbol{K}$ entry-wise leads to*

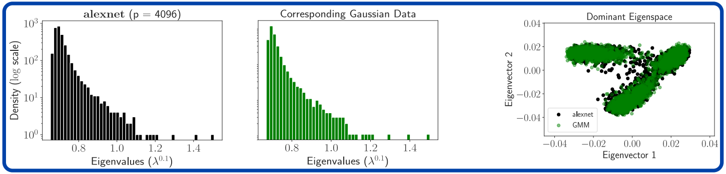
$$\boldsymbol{K}\approx_p f(\tau)\mathbf{1}_n\mathbf{1}_n^\intercal+f'(\tau)\left(\boldsymbol{Z}^\intercal\boldsymbol{Z}/p+\boldsymbol{J}\boldsymbol{\Phi}_{\{\boldsymbol{\mu}_\ell\}_{\ell=1}^k}\boldsymbol{J}^\intercal\right)+f''(\tau)\boldsymbol{J}\boldsymbol{\Phi}_{\{\boldsymbol{\Sigma}_\ell\}_{\ell=1}^k}\boldsymbol{J}^\intercal+*$$

$\boldsymbol{\Phi}_{\{\boldsymbol{\mu}_\ell\}_{\ell=1}^k},\boldsymbol{\Phi}_{\{\boldsymbol{\Sigma}_\ell\}_{\ell=1}^k}$ ***low-rank*** *depending solely on* $\{\boldsymbol{\mu}_\ell,\boldsymbol{\Sigma}_\ell\}_{\ell=1}^k$.

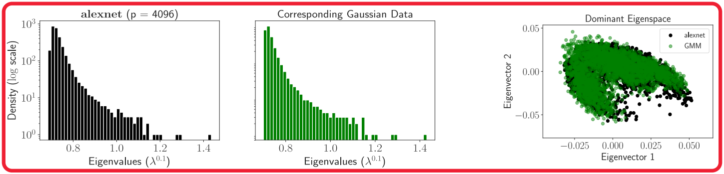
- $\boldsymbol{K}$ behaves as **spiked RMT** model.
- Classification **performance** depends on $f'(\tau)$, $f''(\tau)$, $\{\boldsymbol{\mu}_\ell,\boldsymbol{\Sigma}_\ell\}_{\ell=1}^k$.
- **Universality:** only first and second order statistics matter!

Experiments: Spectrum of Kernel Matrices

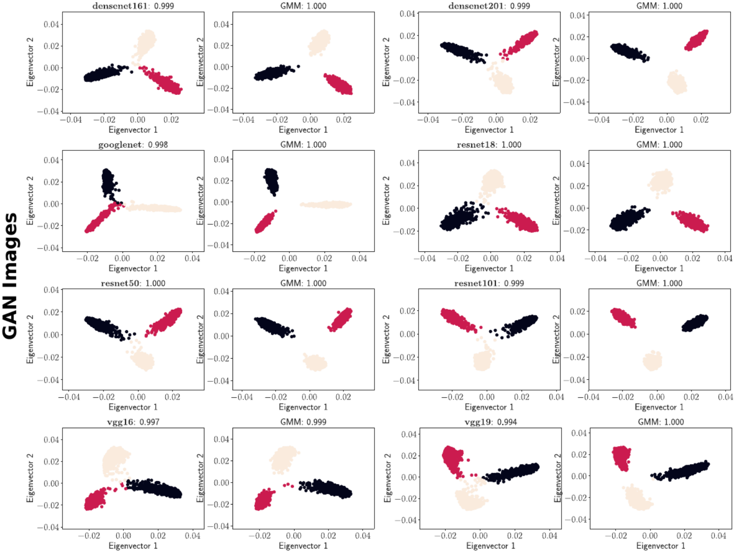
GAN data



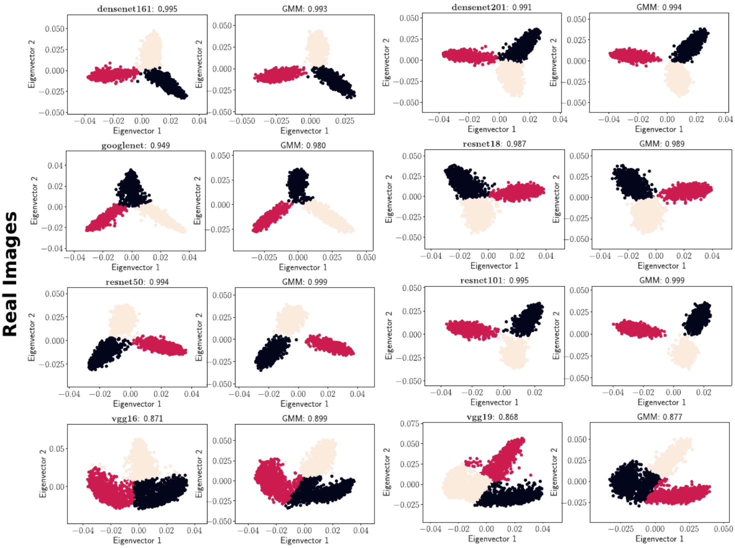
Real data



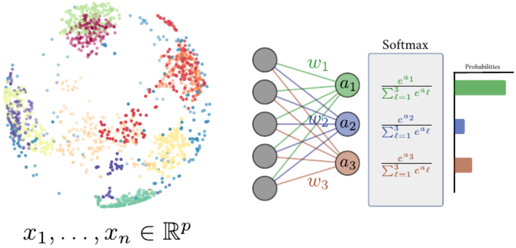
Experiments: Spectral Clustering (k-means: GAN data)



Experiments: Spectral Clustering (k-means: Real data)



Some ML methods under Concentration



Contribution 4

Implicit Classifiers: The Softmax Classifier

MEA. Seddik, *C. Louart, R. Couillet, M. Tamaazousti, "**The Unexpected Deterministic and Universal Behavior of Large Softmax Classifiers**", (submitted to) AISTATS'2021.*

Model & Assumptions

(A1) Data matrix (distributed in k classes $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$):

$$\boldsymbol{X} = \left[\underbrace{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_1}}_{\propto \mathcal{E}_{q_1}}, \underbrace{\boldsymbol{x}_{n_1+1}, \ldots, \boldsymbol{x}_{n_2}}_{\propto \mathcal{E}_{q_2}}, \ldots, \underbrace{\boldsymbol{x}_{n-n_k+1}, \ldots, \boldsymbol{x}_n}_{\propto \mathcal{E}_{q_k}} \right] \in \mathbb{R}^{p \times n}$$

Model statistics: $\boldsymbol{\mu}_\ell = \mathbb{E}_{\boldsymbol{x}_i \in \mathcal{C}_\ell}[\boldsymbol{x}_i], \quad \boldsymbol{\Sigma}_\ell = \mathbb{E}_{\boldsymbol{x}_i \in \mathcal{C}_\ell}[\boldsymbol{x}_i \boldsymbol{x}_i^\intercal] - \boldsymbol{\mu}_\ell \boldsymbol{\mu}_\ell^\intercal$

(A2) Growth rate assumptions: As $p \to \infty$,

1. $p/n \to c \in (0, \infty)$.
2. k fixed.
3. $\|\boldsymbol{\mu}_\ell\| = \mathcal{O}(1)$.

The Softmax classifier: Minimize:

$$\mathcal{L}(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{k} y_{i\ell} \log p_{i\ell} + \frac{1}{2} \sum_{\ell=1}^{k} \lambda_\ell \|\boldsymbol{w}_\ell\|^2$$

$$p_{i\ell} = \frac{\exp(\boldsymbol{w}_\ell^\intercal \boldsymbol{x}_i)}{\sum_{j=1}^{k} \exp(\boldsymbol{w}_j^\intercal \boldsymbol{x}_i)}, \quad \boldsymbol{W} \equiv [\boldsymbol{w}_1^\intercal, \ldots, \boldsymbol{w}_k^\intercal]^\intercal \in \mathbb{R}^{pk}$$

Propagation of the Concentration to Softmax

Implicit equation

$$\nabla_{\boldsymbol{w}_\ell}\mathcal{L} = \boldsymbol{0} \quad \Rightarrow \quad \lambda_\ell \boldsymbol{w}_\ell = -\frac{1}{n}\sum_{i=1}^{n}\left(y_{i\ell} - \frac{\exp(\boldsymbol{w}_\ell^\intercal \boldsymbol{x}_i)}{\sum_{j=1}^{k}\exp(\boldsymbol{w}_j^\intercal \boldsymbol{x}_i)}\right)\boldsymbol{x}_i$$

Equivalently (scalar case for some $f : \mathbb{R} \to \mathbb{R}$)

$$\boldsymbol{w} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{w}^\intercal \boldsymbol{x}_i)\boldsymbol{x}_i \in \mathbb{R}^p \quad \Rightarrow \quad \boxed{\boldsymbol{w} = \Psi(\boldsymbol{w}) \equiv \frac{1}{n}\boldsymbol{X} f(\boldsymbol{X}^\intercal \boldsymbol{w})}$$

Contractivity of Ψ

Ψ is requested to be $(1-\varepsilon)$-Lipschitz for some $\varepsilon > 0$ or equivalently

$$\mathcal{A}_{\boldsymbol{w}} = \left\{ \frac{1}{n}\|f\|_\infty \|\boldsymbol{X}\boldsymbol{X}^\intercal\| \geq 1-\varepsilon \right\} \quad \text{has low probability.}$$

(A3) $\exists \varepsilon > 0$ independent of p, n s.t. $\frac{1}{n}\|f\|_\infty \|\boldsymbol{X}\boldsymbol{X}^\intercal\| \leq 1 - 2\varepsilon$.

Theorem (Concentration of $\boldsymbol{w}$ (SLCT'20))

*Under **(A1-3)**, $\mathbb{P}(\mathcal{A}_{\boldsymbol{w}}) \propto e^{-n}$ and $\boldsymbol{w} \propto \mathcal{E}_q\left(n^{-\frac{1}{2}}\right) \mid \mathcal{A}_{\boldsymbol{w}}$.*

Estimation of the Weights Statistics (SLCT'20)

Let $\mu_w = \mathbb{E}[w] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[f(x_i^\intercal w)x_i]$

Breaking the Weights-data Dependence

1. **Leave-one-data-out:** $w_{-i} = \frac{1}{n} X_{-i} f(X_{-i}^\intercal w_{-i})$
2. **Resolvent matrix:** $Q_{-i} = \left(I_p - \frac{1}{n} X_{-i} D X_{-i}^\intercal\right)^{-1}$ with D diagonal
3. **Link w and w_{-i}:** $x_i^\intercal w \approx x_i^\intercal w_{-i} + \frac{1}{n} x_i^\intercal Q_{-i} x_i f(x_i^\intercal w)$
4. $Q_{-i} \leftrightarrow \bar{Q}$: so $\frac{1}{n} x_i^\intercal Q_{-i} x_i \to \delta_\ell = \frac{1}{n}\operatorname{tr}(\Sigma_\ell \bar{Q})$
5. **Hence:** $f(x_i^\intercal w) \approx f(x_i^\intercal w_{-i} + \delta_\ell f(x_i^\intercal w)) = g_\ell(x_i^\intercal w_{-i})$

Stein's Lemma

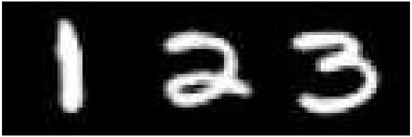
1. **Gaussianity** of $z_i = x_i^\intercal w_{-i}$
2. $\mathbb{E}[f(x_i^\intercal w) x_i] \approx \mathbb{E}[g_\ell(x_i^\intercal w_{-i}) x_i] \approx \mathbb{E}[g_\ell(z_i)]\mu_\ell + \mathbb{E}[g_\ell'(z_i)]\Sigma_\ell \mu_w$

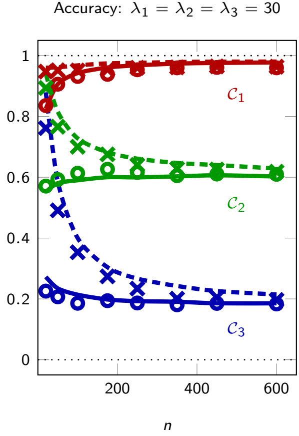
Similarly with $\Sigma_w = \mathbb{E}[ww^\intercal] - \mu_w \mu_w^\intercal$

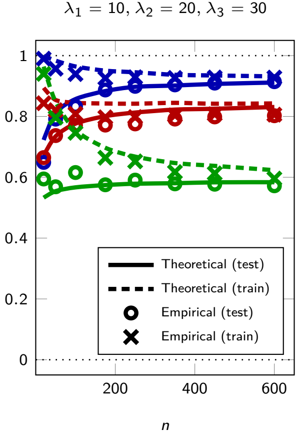
$$\Rightarrow \quad (\mu_w, \Sigma_w) = \Psi_{\{\mu_\ell, \Sigma_\ell\}_{\ell=1}^k}(\mu_w, \Sigma_w)$$

Universality: only first and second order statistics matter!

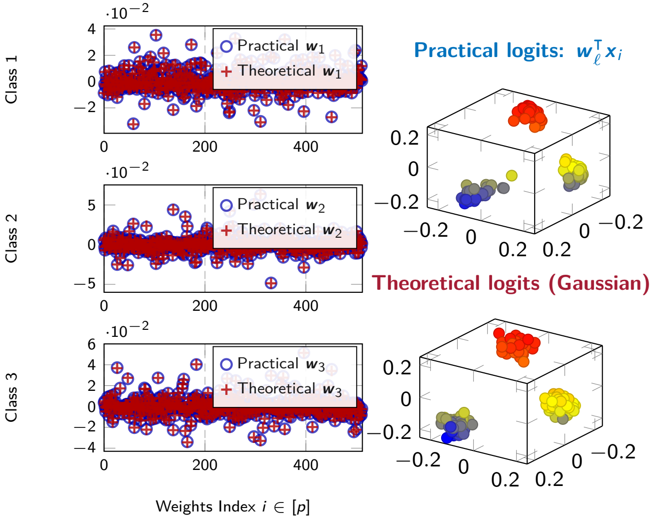
Simulations with MNIST Generated Data



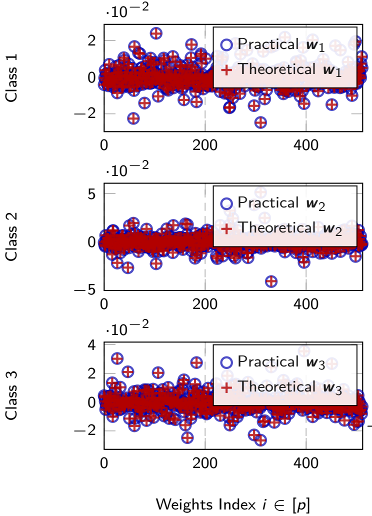




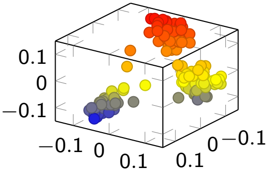
Experimental validation (GAN data)



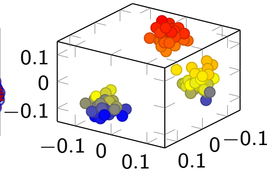
Experimental validation (Real data)



Practical logits: $\boldsymbol{w}_\ell^\top \boldsymbol{x}_i$



Theoretical logits (Gaussian)



Outline

Conclusions & Perspectives

Conclusions

- **Concentrated Vectors** are **very likely** appropriate for realistic data modelling.
- RMT can ***anticipate*** performances of ML classifiers for **Concentrated Vectors** . . . so for **realistic** data (so far GAN data).
- ***Universality*** of ML classifiers regardless of data distribution.

Perspectives

- Study of **non-convex** (e.g., deep neural nets) optimization problems. Learning of two layers networks **(Goldt+'20)**. What statistics encoded by hidden layers?
- More to be explored with RMT: active and reinforcement learning, generative models, graph-based methods (GNNs), . . . etc.
- Generalize these ideas to **other modalities** (NLP?). For NLP, with RNNs $\boldsymbol{z}_t = \mathrm{RNN}(\boldsymbol{z}_{t-1}, \boldsymbol{w}_{t-1})$ with $\boldsymbol{z}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$? Word embeddings seem to concentrate **(Couillet+'20)**.

Thank you for your attention!

Publications

1. **MEA.Seddik**, C.Louart, R.Couillet, M.Tamaazousti, "*The Unexpected Deterministic and Universal Behavior of Large Softmax Classifiers*", (submitted to) **AISTATS'20**.
2. **MEA.Seddik**, R.Couillet, M.Tamaazousti, "*A Random Matrix Analysis of Learning with alpha-Dropout*", **ICML'20** Artemiss Workshop.
3. **MEA.Seddik**, C.Louart, M.Tamaazousti, R.Couillet, "*Random Matrix Theory Proves that Deep Learning Representations of GAN-data Behave as Gaussian Mixtures*", **ICML'20**.
4. **MEA.Seddik**, M.Tamaazousti, R.Couillet, "*Why do Random Matrices Explain Learning? An Argument of Universality Offered by GANs*", **GRETSI'19**.
5. **MEA.Seddik**, M.Tamaazousti, R.Couillet, "*Kernel Random Matrices of Large Concentrated Data: The Example of GAN-generated Images*", **ICASSP'19**.
6. **MEA.Seddik**, M.Tamaazousti, R.Couillet, "*A Kernel Random Matrix-Based Approach for Sparse PCA*", **ICLR'19**.

Other Contributions

1. **MEA.Seddik**, H.Essafi, A.Benzine, M.Tamaazousti, "*Lightweight Neural Networks from PCA LDA Based Distilled Dense Neural Networks*", **ICIP'20**.
2. **MEA.Seddik**, M.Tamaazousti, J.Lin, "*Generative Collaborative Networks for Single Image SuperResolution*", **Neurocomputing'19**.

$+5$ **patents.**

References I

Vladimir A Marvcenko and Leonid Andreevich Pastur. "Distribution of eigenvalues for some sets of random matrices". In: *Mathematics of the USSR-Sbornik* 1.4 (1967), p. 457

Noureddine El Karoui et al. "The spectrum of kernel random matrices". In: *The Annals of Statistics* 38.1 (2010), pp. 1–50

Romain Couillet, Florent Benaych-Georges, et al. "Kernel spectral clustering of large dimensional data". In: *Electronic Journal of Statistics* 10.1 (2016), pp. 1393–1454

Romain Couillet et al. "Word Representations Concentrate and This is Good News!" In: (2020)

Sebastian Goldt et al. "The gaussian equivalence of generative models for learning with two-layer neural networks". In: *arXiv preprint arXiv:2006.14709* (2020)

Zhenyu Liao. "Random Matrix Advances in Large Dimensional Machine Learning". In: (2019)

Michel Ledoux. *The concentration of measure phenomenon*. 89. American Mathematical Soc., 2005

Xiaoyi Mai. "Méthodes des matrices aléatoires pour l'apprentissage en grandes dimensions". PhD thesis. Paris Saclay, 2019

Jeffrey Pennington and Yasaman Bahri. "Geometry of neural network loss surfaces via random matrix theory". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 2798–2806

References II

Jeffrey Pennington and Pratik Worah. "Nonlinear random matrix theory for deep learning". In: *Advances in Neural Information Processing Systems*. 2017, pp. 2637–2646

Anna Choromanska et al. "The loss surfaces of multilayer networks". In: *Artificial intelligence and statistics*. 2015, pp. 192–204

Noureddine El Karoui et al. "On robust regression with high-dimensional predictors". In: *Proceedings of the National Academy of Sciences* 110.36 (2013), pp. 14557–14562

Malik Tiomoko, Cosme Louart, and Romain Couillet. "Large Dimensional Asymptotics of Multi-Task Learning". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 8787–8791

Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems*. 2014, pp. 2672–2680

Andrew Brock, Jeff Donahue, and Karen Simonyan. "Large scale gan training for high fidelity natural image synthesis". In: *arXiv preprint arXiv:1809.11096* (2018)

Cosme Louart, Zhenyu Liao, Romain Couillet, et al. "A random matrix approach to neural networks". In: *The Annals of Applied Probability* 28.2 (2018), pp. 1190–1248

Cosme Louart and Romain Couillet. "Concentration of measure and large random matrices with an application to sample covariance matrices". In: *arXiv preprint arXiv:1805.08295* (2018)

References III

Mohamed El Amine Seddik et al. "Random matrix theory proves that deep learning representations of gan-data behave as gaussian mixtures". In: *arXiv preprint arXiv:2001.08370* (2020)

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. URL: https://openreview.net/forum?id=rkgBHoCqYX

Mohamed El Amine Seddik, Mohamed Tamaazousti, and Romain Couillet. "Pourquoi les matrices aléatoires expliquent l'apprentissage ? Un argument d'universalité offert par les GANs". In: *Colloque francophone de traitement du signal et des images*. 2019

Mohamed El Amine Seddik, Mohamed Tamaazousti, and John Lin. "Generative Collaborative Networks for Single Image Super-Resolution". In: *arXiv:1902.10467* (2019)

Mohamed El Amine Seddik, Mohamed Tamaazousti, and Romain Couillet. "Kernel Random Matrices of Large Concentrated Data: The Example of GAN-Generated Images". In: *International Conference on Acoustics, Speech, and Signal Processing*. 2019

Mohamed El Amine Seddik et al. "Lightweight Neural Networks from PCA LDA Based Distilled Dense Neural Networks". In: *International Conference on Image Processing*. 2020

Youssef Tamaazousti et al. "Learning More Universal Representations for Transfer-Learning". In: *arXiv:1712.09708* (2019)