Random Matrix Theory for Big-Data and Machine Learning

Mohamed El Amine Seddik
-
Supervised by:
Mohamed Tamaazousti and Romain Couillet

CEA List & Centralesupélec

ICTP Trieste - 10 December 2018

Outline

Outline

Introduction

In machine learning,

Introduction

In machine learning,

- We are given data

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$$

Introduction

In machine learning,

- We are given data

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$$

- We aim at performing

Regression, Classification, Clustering etc.

Introduction

In machine learning,

- We are given data

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$$

- We aim at performing

Regression, Classification, Clustering etc.

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

Introduction

In machine learning,

- We are given data

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$$

- We aim at performing

Regression, Classification, Clustering etc.

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

Example: $\mathbf{x}_i^\intercal \mathbf{x}_j$

Introduction

In machine learning,

- We are given data

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$$

- We aim at performing

Regression, Classification, Clustering etc.

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

Example: $\mathbf{x}_i^\intercal \mathbf{x}_j$

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

Introduction

In machine learning,

- We are given data

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$$

- We aim at performing

Regression, Classification, Clustering etc.

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

Example: $\mathbf{x}_i^\intercal \mathbf{x}_j$

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

- **How does it behave?**

Outline

Behavior of Large Random Gram Matrices

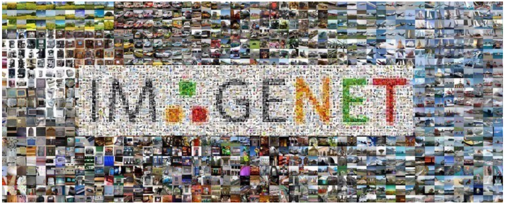
- The behavior of $\mathbf{X}^\intercal\mathbf{X}$ "depends" on p and n.

Behavior of Large Random Gram Matrices

- The behavior of $\mathbf{X}^\intercal \mathbf{X}$ "depends" on p and n.
- In ML, we generally face **large and numerous data**.

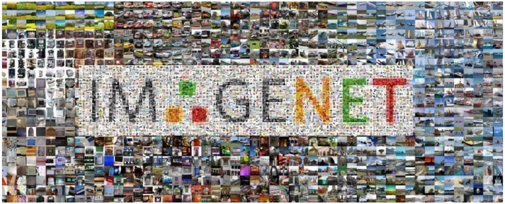
Behavior of Large Random Gram Matrices

- The behavior of $\mathbf{X}^\mathsf{T}\mathbf{X}$ "depends" on p and n.
- In ML, we generally face **large and numerous data**.



Behavior of Large Random Gram Matrices

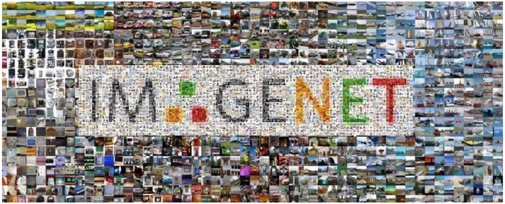
- The behavior of $\mathbf{X}^\intercal \mathbf{X}$ "depends" on p and n.
- In ML, we generally face **large and numerous data**.



Dimension: **150,528**

Behavior of Large Random Gram Matrices

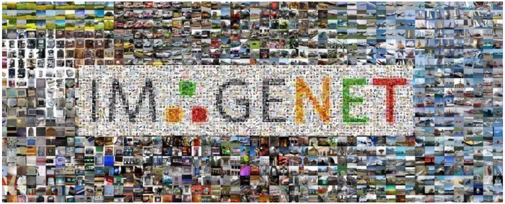
- The behavior of $\mathbf{X}^\intercal \mathbf{X}$ "depends" on p and n.
- In ML, we generally face **large and numerous data**.



Dimension: **150,528**
N° of images: **14,197,122**

Behavior of Large Random Gram Matrices

- The behavior of $\mathbf{X}^\intercal \mathbf{X}$ "depends" on p and n.
- In ML, we generally face **large and numerous data**.



Dimension: **150,528**
N° of images: **14,197,122**

- Which is the **Random Matrix Theory Regime**

$$\boxed{p, n \to \infty, \text{ with } p/n \to c \in (0, \infty)}$$

Outline

The Marčenko–Pastur law [Marčenko,Pastur'67]

- If we assume all $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I}_p)$

The Marčenko–Pastur law [Marčenko,Pastur'67]

- If we assume all $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I}_p)$

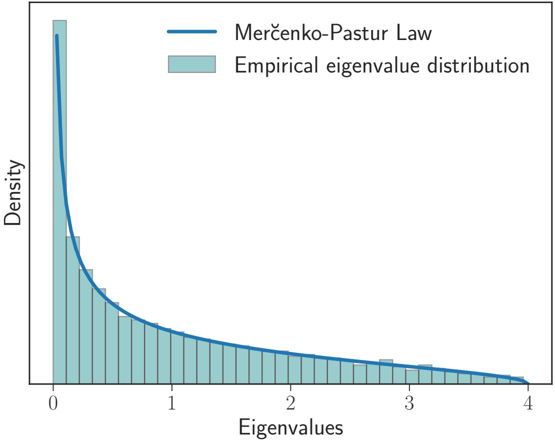


Figure 1: Histogram of the eigenvalues of $\frac{1}{p}\mathbf{X}^\intercal\mathbf{X}$ for $n = p = 1000$.

The Marčenko–Pastur law [Marčenko,Pastur'67]

Definition (Empirical Spectral Density)

Empirical Spectral Density (e.s.d.) μ_n of Hermitian matrix $\mathbf{A}_n \in \mathbb{R}^{n \times n}$ is $\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i(\mathbf{A}_n)}$.

The Marčenko–Pastur law [Marčenko,Pastur'67]

Definition (Empirical Spectral Density)

Empirical Spectral Density (e.s.d.) μ_n of Hermitian matrix $\mathbf{A}_n \in \mathbb{R}^{n\times n}$ is $\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(\mathbf{A}_n)}$.

Theorem (Marčenko–Pastur law)

$\mathbf{X} \in \mathbb{R}^{p\times n}$ *with i.i.d. zero mean, unit variance entries.*
As $p, n \to \infty$ *with* $n/p \to c \in (0,\infty)$, *e.s.d.* μ_n *of* $\frac{1}{p}\mathbf{X}^\mathsf{T}\mathbf{X}$ *satisfies*

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

weakly, where μ_c *has continuous density* f_c *with compact support* $[\lambda^-, \lambda^+] = [(1-\sqrt{c})^2, (1+\sqrt{c})^2]$

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

The Marčenko–Pastur law

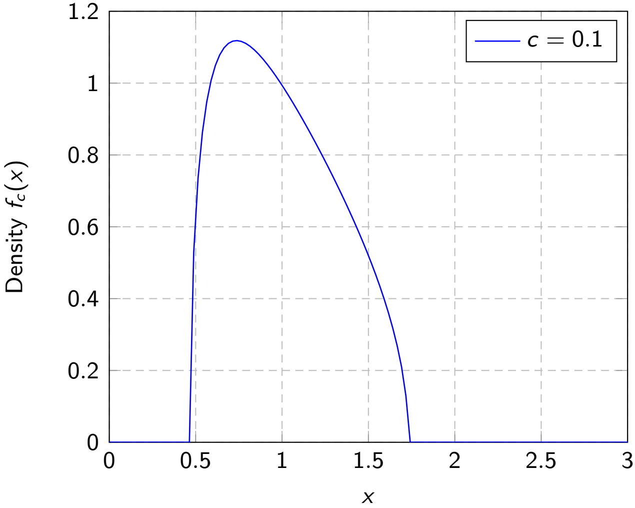


Figure 2: Marčenko-Pastur law for different limit ratios $c = \lim_{p\to\infty} p/n$.

The Marčenko–Pastur law

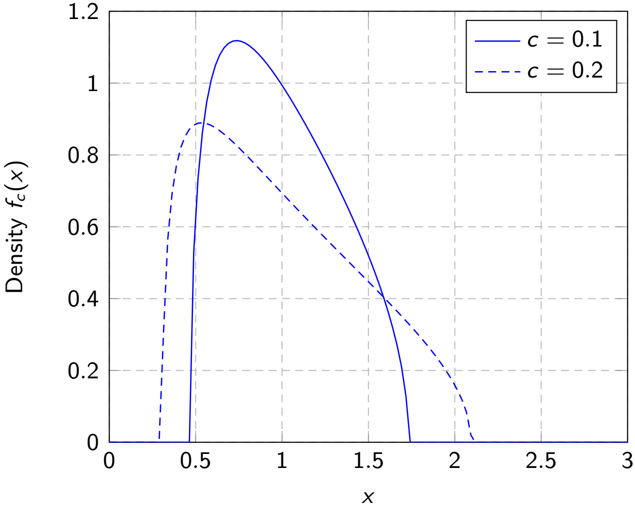


Figure 2: Marčenko-Pastur law for different limit ratios $c = \lim_{p\to\infty} p/n$.

The Marčenko–Pastur law

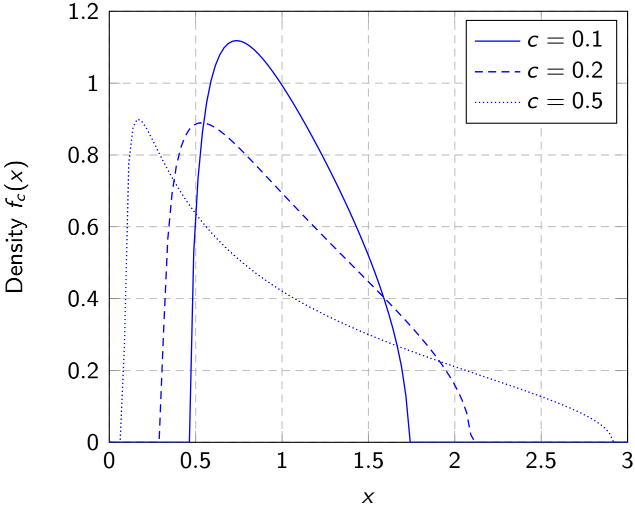


Figure 2: Marčenko-Pastur law for different limit ratios $c = \lim_{p\to\infty} p/n$.

Outline

Gaussian Mixtures (Spiked Model)

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

Gaussian Mixtures (Spiked Model)

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

$$\mathbf{X} = [\underbrace{\mathbf{x}_1, \ldots, \mathbf{x}_{\frac{n}{2}}}_{\sim\mathcal{N}(+\mathbf{m},\mathbf{I}_p)}, \underbrace{\mathbf{x}_{\frac{n}{2}+1}, \ldots, \mathbf{x}_n}_{\sim\mathcal{N}(-\mathbf{m},\mathbf{I}_p)}]$$

Gaussian Mixtures (Spiked Model)

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

$$\mathbf{X} = [\underbrace{\mathbf{x}_1, \ldots, \mathbf{x}_{\frac{n}{2}}}_{\sim\mathcal{N}(+\mathbf{m},\mathbf{I}_p)}, \underbrace{\mathbf{x}_{\frac{n}{2}+1}, \ldots, \mathbf{x}_n}_{\sim\mathcal{N}(-\mathbf{m},\mathbf{I}_p)}]$$

- Which can be written as

$$\mathbf{X} = \mathbf{m}\,\mathbf{y}^\mathsf{T} + \mathbf{Z}$$

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

Gaussian Mixtures (Spiked Model)

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

$$\mathbf{X} = [\underbrace{\mathbf{x}_1, \ldots, \mathbf{x}_{\frac{n}{2}}}_{\sim\mathcal{N}(+\mathbf{m}, \mathbf{I}_p)}, \underbrace{\mathbf{x}_{\frac{n}{2}+1}, \ldots, \mathbf{x}_n}_{\sim\mathcal{N}(-\mathbf{m}, \mathbf{I}_p)}]$$

- Which can be written as

$$\mathbf{X} = \mathbf{m}\,\mathbf{y}^\intercal + \mathbf{Z}$$

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

- We have then

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

Gaussian Mixtures (Spiked Model)

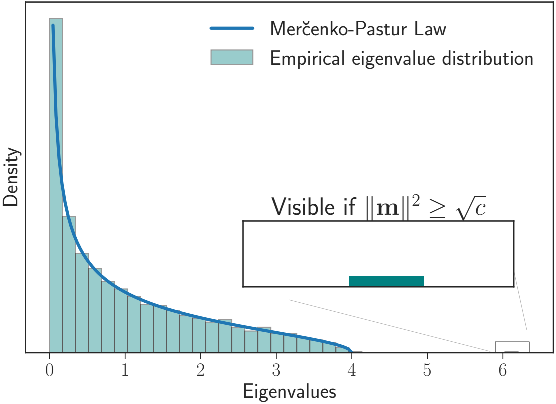


Figure 3: Histogram of the eigenvalues of $\frac{1}{p}\mathbf{X}^\intercal\mathbf{X}$ for $n = p = 1000$.

Gaussian Mixtures (Spiked Model)

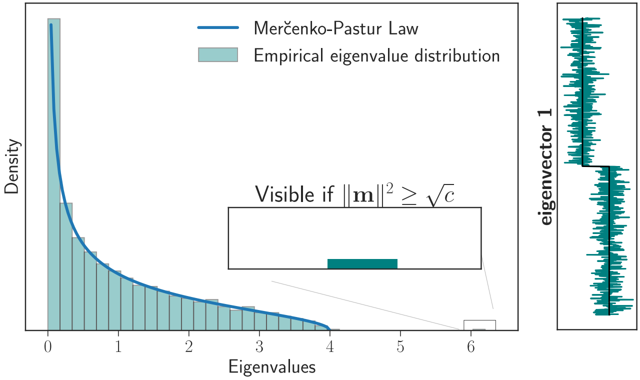


Figure 4: Histogram of the eigenvalues of $\frac{1}{p}\mathbf{X}^\intercal\mathbf{X}$ and its dominant eigenvector for $n = p = 1000$.

Spiked Models

Theorem (Eigenvalues [Baik, Silverstein'06])

Let

- $\mathbf{Z}$ *with i.i.d. zero mean, unit variance,* $\mathbb{E}|Z_{ij}|^4 < \infty$
- $\mathbf{X} = \mathbf{m}\mathbf{y}^\intercal + \mathbf{Z}$

Spiked Models

Theorem (Eigenvalues [Baik, Silverstein'06])

Let

- $\mathbf{Z}$ *with i.i.d. zero mean, unit variance,* $\mathbb{E}|Z_{ij}|^4 < \infty$
- $\mathbf{X} = \mathbf{m}\mathbf{y}^\mathsf{T} + \mathbf{Z}$

Then, under the RMT regime, if $\|\mathbf{m}\|^2 > \sqrt{c}$

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

Spiked Models

Theorem (Eigenvectors [Paul'07])

Let

- $\mathbf{Z}$ *with i.i.d. zero mean, unit variance,* $\mathbb{E}|Z_{ij}|^4 < \infty$
- $\mathbf{X} = \mathbf{m}\mathbf{y}^\mathsf{T} + \mathbf{Z}$

Spiked Models

Theorem (Eigenvectors [Paul'07])

Let

- $\mathbf{Z}$ *with i.i.d. zero mean, unit variance,* $\mathbb{E}|Z_{ij}|^4 < \infty$
- $\mathbf{X} = \mathbf{m}\mathbf{y}^\mathsf{T} + \mathbf{Z}$

Then, under the RMT regime, for $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$ *deterministic and* $\hat{\mathbf{u}}$ *eigenvector corresponding to* $\lambda_{\max}\left(\frac{1}{p}\mathbf{X}^\mathsf{T}\mathbf{X}\right)$,

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

In particular,

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

Spiked Models

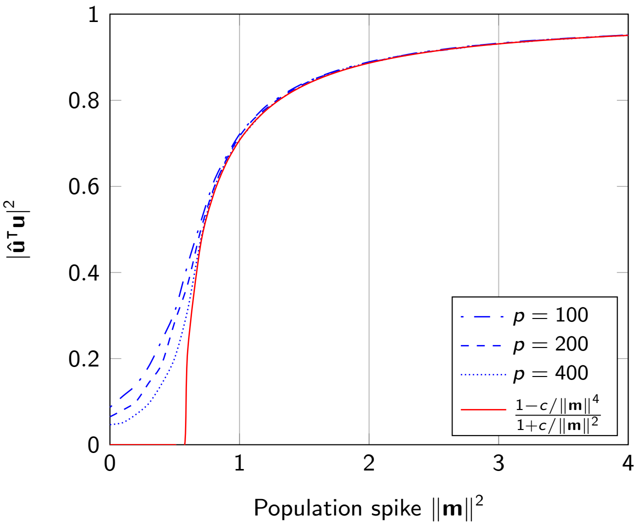


Figure 5: Simulated versus limiting $|\hat{\mathbf{u}}^\mathsf{T}\mathbf{u}|^2$, $p/n = 1/3$, varying $\|\mathbf{m}\|^2$.

Outline

Outline

Kernel Spectral Clustering

Problem Statement

- Given data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$
- Objective: "cluster" in k **similarity** classes.

Kernel Spectral Clustering

Problem Statement

- Given data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$
- Objective: "cluster" in k **similarity** classes.
- Based on a kernel matrix

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

Kernel Spectral Clustering

Problem Statement

- Given data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$
- Objective: "cluster" in k **similarity** classes.
- Based on a kernel matrix

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

Intuition (from small dimensions)

$$\mathbf{K} = \begin{pmatrix} \gg 1 & \ll 1 & \ll 1 \\ \ll 1 & \gg 1 & \ll 1 \\ \ll 1 & \ll 1 & \gg 1 \end{pmatrix} \begin{matrix} \updownarrow \mathcal{C}_1 \\ \updownarrow \mathcal{C}_2 \\ \updownarrow \mathcal{C}_3 \end{matrix}$$

Kernel Spectral Clustering

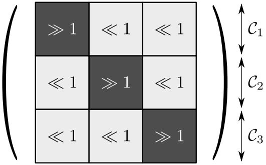
Problem Statement

- Given data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$
- Objective: "cluster" in k **similarity** classes.
- Based on a kernel matrix

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

Intuition (from small dimensions)

$\mathbf{K} =$



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

Outline

Gaussian Mixture Model and Growth Assumptions

Consider data distributed in k classes as

$$\mathbf{X} = [\underbrace{\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}}_{\sim\mathcal{N}(\mathbf{m_1}, \mathbf{C}_1)}, \underbrace{\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_{n_2}}_{\sim\mathcal{N}(\mathbf{m_2}, \mathbf{C}_2)}, \ldots, \underbrace{\mathbf{x}_{n-n_k+1}, \ldots, \mathbf{x}_n}_{\sim\mathcal{N}(\mathbf{m_k}, \mathbf{C}_k)}]$$

Gaussian Mixture Model and Growth Assumptions

Consider data distributed in k classes as

$$\mathbf{X} = [\underbrace{\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}}_{\sim\mathcal{N}(\mathbf{m_1}, \mathbf{C}_1)}, \underbrace{\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_{n_2}}_{\sim\mathcal{N}(\mathbf{m_2}, \mathbf{C}_2)}, \ldots, \underbrace{\mathbf{x}_{n-n_k+1}, \ldots, \mathbf{x}_n}_{\sim\mathcal{N}(\mathbf{m_k}, \mathbf{C}_k)}]$$

And let

$$\mathbf{m} = \sum_{\ell=1}^{k} \frac{n_\ell}{n}\mathbf{m}_\ell,\ \bar{\mathbf{m}}_\ell = \mathbf{m}_\ell - \mathbf{m},\quad \mathbf{C} = \sum_{\ell=1}^{k} \frac{n_\ell}{n}\mathbf{C}_\ell,\ \bar{\mathbf{C}}_\ell = \mathbf{C}_\ell - \mathbf{C}$$

Gaussian Mixture Model and Growth Assumptions

Consider data distributed in k classes as

$$\mathbf{X} = [\underbrace{\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}}_{\sim\mathcal{N}(\mathbf{m_1}, \mathbf{C}_1)}, \underbrace{\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_{n_2}}_{\sim\mathcal{N}(\mathbf{m_2}, \mathbf{C}_2)}, \ldots, \underbrace{\mathbf{x}_{n-n_k+1}, \ldots, \mathbf{x}_n}_{\sim\mathcal{N}(\mathbf{m_k}, \mathbf{C}_k)}]$$

And let

$$\mathbf{m} = \sum_{\ell=1}^{k} \frac{n_\ell}{n}\mathbf{m}_\ell, \; \bar{\mathbf{m}}_\ell = \mathbf{m}_\ell - \mathbf{m}, \;\; \mathbf{C} = \sum_{\ell=1}^{k} \frac{n_\ell}{n}\mathbf{C}_\ell, \; \bar{\mathbf{C}}_\ell = \mathbf{C}_\ell - \mathbf{C}$$

Assumption (Growth rate)

As $n \to \infty$,

Gaussian Mixture Model and Growth Assumptions

Consider data distributed in k classes as

$$\mathbf{X} = [\underbrace{\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}}_{\sim\mathcal{N}(\mathbf{m_1}, \mathbf{C}_1)}, \underbrace{\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_{n_2}}_{\sim\mathcal{N}(\mathbf{m_2}, \mathbf{C}_2)}, \ldots, \underbrace{\mathbf{x}_{n-n_k+1}, \ldots, \mathbf{x}_n}_{\sim\mathcal{N}(\mathbf{m_k}, \mathbf{C}_k)}]$$

And let

$$\mathbf{m} = \sum_{\ell=1}^{k} \frac{n_\ell}{n}\mathbf{m}_\ell,\ \bar{\mathbf{m}}_\ell = \mathbf{m}_\ell - \mathbf{m},\quad \mathbf{C} = \sum_{\ell=1}^{k} \frac{n_\ell}{n}\mathbf{C}_\ell,\ \bar{\mathbf{C}}_\ell = \mathbf{C}_\ell - \mathbf{C}$$

Assumption (Growth rate)

As $n \to \infty$,

1. **Data:** $\frac{p}{n} \to c_0 \in (0, \infty)$, $\frac{n_\ell}{n} \to c_\ell \in (0, 1)$

Gaussian Mixture Model and Growth Assumptions

Consider data distributed in k classes as

$$\mathbf{X} = [\underbrace{\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}}_{\sim\mathcal{N}(\mathbf{m_1}, \mathbf{C}_1)}, \underbrace{\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_{n_2}}_{\sim\mathcal{N}(\mathbf{m_2}, \mathbf{C}_2)}, \ldots, \underbrace{\mathbf{x}_{n-n_k+1}, \ldots, \mathbf{x}_n}_{\sim\mathcal{N}(\mathbf{m_k}, \mathbf{C}_k)}]$$

And let

$$\mathbf{m} = \sum_{\ell=1}^{k} \frac{n_\ell}{n} \mathbf{m}_\ell, \; \bar{\mathbf{m}}_\ell = \mathbf{m}_\ell - \mathbf{m}, \quad \mathbf{C} = \sum_{\ell=1}^{k} \frac{n_\ell}{n} \mathbf{C}_\ell, \; \bar{\mathbf{C}}_\ell = \mathbf{C}_\ell - \mathbf{C}$$

Assumption (Growth rate)

As $n \to \infty$,

1. **Data:** $\frac{p}{n} \to c_0 \in (0, \infty)$, $\frac{n_\ell}{n} \to c_\ell \in (0, 1)$
2. **Mean:** $\|\bar{\mathbf{m}}_\ell\| = \mathcal{O}(1)$

Gaussian Mixture Model and Growth Assumptions

Consider data distributed in k classes as

$$\mathbf{X} = [\underbrace{\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}}_{\sim\mathcal{N}(\mathbf{m_1}, \mathbf{C}_1)}, \underbrace{\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_{n_2}}_{\sim\mathcal{N}(\mathbf{m_2}, \mathbf{C}_2)}, \ldots, \underbrace{\mathbf{x}_{n-n_k+1}, \ldots, \mathbf{x}_n}_{\sim\mathcal{N}(\mathbf{m_k}, \mathbf{C}_k)}]$$

And let

$$\mathbf{m} = \sum_{\ell=1}^{k} \frac{n_\ell}{n} \mathbf{m}_\ell,\ \bar{\mathbf{m}}_\ell = \mathbf{m}_\ell - \mathbf{m},\quad \mathbf{C} = \sum_{\ell=1}^{k} \frac{n_\ell}{n} \mathbf{C}_\ell,\ \bar{\mathbf{C}}_\ell = \mathbf{C}_\ell - \mathbf{C}$$

Assumption (Growth rate)

As $n \to \infty$,

1. **Data:** $\frac{p}{n} \to c_0 \in (0, \infty)$, $\frac{n_\ell}{n} \to c_\ell \in (0, 1)$
2. **Mean:** $\|\bar{\mathbf{m}}_\ell\| = \mathcal{O}(1)$
3. **Covariance:** $\|\bar{\mathbf{C}}_\ell\| = \mathcal{O}(1)$, $\operatorname{tr}\bar{\mathbf{C}}_\ell = \mathcal{O}(\sqrt{p})$, $\operatorname{tr}\bar{\mathbf{C}}_a\bar{\mathbf{C}}_b = \mathcal{O}(p)$

Small Dimension vs High Dimension!

$$\mathbf{K} = \begin{pmatrix} \gg 1 & \ll 1 & \ll 1 \\ \ll 1 & \gg 1 & \ll 1 \\ \ll 1 & \ll 1 & \gg 1 \end{pmatrix} \begin{matrix} \updownarrow \mathcal{C}_1 \\ \updownarrow \mathcal{C}_2 \\ \updownarrow \mathcal{C}_3 \end{matrix}$$

Small Dimension vs High Dimension!

$$\mathbf{K} = \begin{pmatrix} \gg 1 & \ll 1 & \ll 1 \\ \ll 1 & \gg 1 & \ll 1 \\ \ll 1 & \ll 1 & \gg 1 \end{pmatrix} \begin{matrix} \mathcal{C}_1 \\ \mathcal{C}_2 \\ \mathcal{C}_3 \end{matrix} \qquad \mathbf{K} = \begin{pmatrix} \gg 1 & \gg 1 & \gg 1 \\ \gg 1 & \gg 1 & \gg 1 \\ \gg 1 & \gg 1 & \gg 1 \end{pmatrix} \begin{matrix} \mathcal{C}_1 \\ \mathcal{C}_2 \\ \mathcal{C}_3 \end{matrix}$$

Outline

Random Matrix Equivalent [Couillet, Benaych'15]

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

$$\boxed{\max_{1\leq i\neq j\leq n}\left\{\left|\frac{1}{p}\|\mathbf{x}_i-\mathbf{x}_j\|^2-\tau\right|\right\}\xrightarrow{a.s.}0}$$

where $\tau=\frac{2}{p}\operatorname{tr}\mathbf{C}$.

Random Matrix Equivalent [Couillet, Benaych'15]

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

$$\boxed{\max_{1\leq i\neq j\leq n}\left\{\left|\frac{1}{p}\|\mathbf{x}_i-\mathbf{x}_j\|^2-\tau\right|\right\}\xrightarrow{a.s.}0}$$

where $\tau=\frac{2}{p}\mathrm{tr}\mathbf{C}$.

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

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

Random Matrix Equivalent [Couillet, Benaych'15]

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

$$\boxed{\max_{1\leq i\neq j\leq n}\left\{\left|\frac{1}{p}\|\mathbf{x}_i-\mathbf{x}_j\|^2-\tau\right|\right\}\xrightarrow{a.s.}0}$$

where $\tau=\frac{2}{p}\mathrm{tr}\mathbf{C}$.

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

Random Matrix Equivalent [Couillet, Benaych'15]

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

$$\boxed{\max_{1\leq i\neq j\leq n}\left\{\left|\frac{1}{p}\|\mathbf{x}_i-\mathbf{x}_j\|^2-\tau\right|\right\}\xrightarrow{a.s.}0}$$

where $\tau=\frac{2}{p}\mathrm{tr}\mathbf{C}$.

- 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

$$\mathbf{J}=[\mathbf{j}_1,\ldots,\mathbf{j}_k],$$

Random Matrix Equivalent [Couillet, Benaych'15]

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

$$\max_{1\le i\neq j\le n}\left\{\left|\frac{1}{p}\|\mathbf{x}_i-\mathbf{x}_j\|^2-\tau\right|\right\}\xrightarrow{a.s.}0$$

where $\tau=\frac{2}{p}\mathrm{tr}\mathbf{C}$.

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

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

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

$$\mathbf{J}=[\mathbf{j}_1,\ldots,\mathbf{j}_k],\ \mathbf{M}=[\bar{\mathbf{m}}_1,\ldots,\bar{\mathbf{m}}_k]$$

Random Matrix Equivalent [Couillet, Benaych'15]

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

$$\max_{1\leq i\neq j\leq n}\left\{\left|\frac{1}{p}\|\mathbf{x}_i-\mathbf{x}_j\|^2-\tau\right|\right\}\xrightarrow{a.s.}0$$

where $\tau=\frac{2}{p}\mathrm{tr}\mathbf{C}$.

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

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

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

$$\mathbf{J}=[\mathbf{j}_1,\ldots,\mathbf{j}_k],\ \ \mathbf{M}=[\bar{\mathbf{m}}_1,\ldots,\bar{\mathbf{m}}_k]$$

$$\mathbf{t}=\left\{\frac{\mathrm{tr}\bar{\mathbf{C}}_\ell}{\sqrt{p}}\right\}_{\ell=1}^{k},$$

Random Matrix Equivalent [Couillet, Benaych'15]

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

$$\max_{1\leq i\neq j\leq n} \left\{ \left| \frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 - \tau \right| \right\} \xrightarrow{a.s.} 0$$

where $\tau = \frac{2}{p}\mathrm{tr}\mathbf{C}$.

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

$$\mathbf{K} \propto \underbrace{\mathbf{J}\mathbf{A}\mathbf{J}^\mathsf{T}}_{Information} + \underbrace{f'(\tau)\mathbf{Z}^\mathsf{T}\mathbf{Z} + *}_{Noise}$$

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

$$\mathbf{J} = [\mathbf{j}_1, \ldots, \mathbf{j}_k], \ \mathbf{M} = [\bar{\mathbf{m}}_1, \ldots, \bar{\mathbf{m}}_k]$$

$$\mathbf{t} = \left\{ \frac{\mathrm{tr}\bar{\mathbf{C}}_\ell}{\sqrt{p}} \right\}_{\ell=1}^k, \ \mathbf{T} = \left\{ \frac{\mathrm{tr}\bar{\mathbf{C}}_a\bar{\mathbf{C}}_b}{p} \right\}_{a,b=1}^k$$

Theory (with Gaussian assum.) versus MNIST

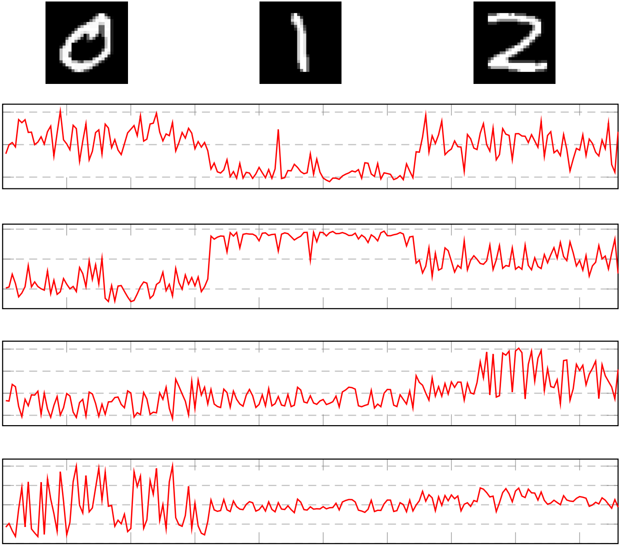


Figure 6: Leading four eigenvectors of $\mathbf{K}$ for MNIST data (**red**) and theoretical findings (**blue**).

Theory (with Gaussian assum.) versus MNIST

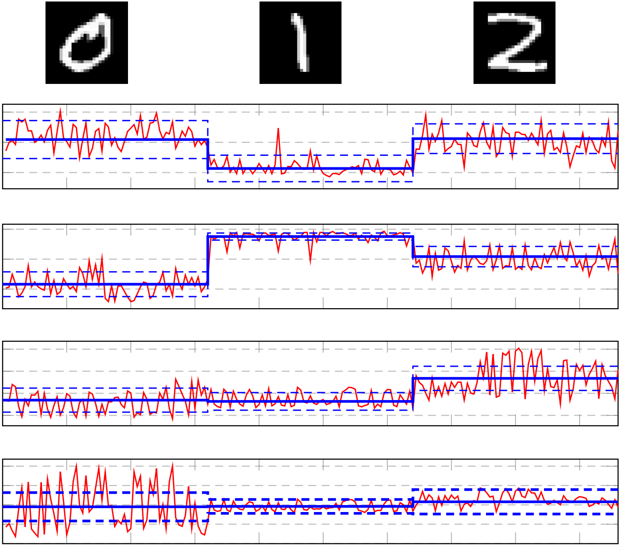


Figure 6: Leading four eigenvectors of $\mathbf{K}$ for MNIST data (**red**) and theoretical findings (**blue**).

Outline

Outline

Why does it work on real data?

- Real data are **unlikely close** to Gaussian.

Why does it work on real data?

- Real data are **unlikely close** to Gaussian.
- However, using **GANs**, it is possible to

$$\text{Real data} \approx \underbrace{\mathcal{F}_1 \circ \mathcal{F}_2 \cdots \circ \mathcal{F}_N}_{\textit{Lipschitz}\text{ operations}}(\text{Gaussian})$$

Why does it work on real data?

- Real data are **unlikely close** to Gaussian.
- However, using **GANs**, it is possible to

$$\text{Real data} \approx \underbrace{\mathcal{F}_1 \circ \mathcal{F}_2 \cdots \circ \mathcal{F}_N}_{\textit{Lipschitz}\text{ operations}}(\text{Gaussian})$$

- On the other hand, a **concentrated** vector $\mathbf{z}$ is defined as For any *Lipschitz* observation $\mathcal{F} : \mathbb{R}^p \to \mathbb{R}$,

$$\mathbb{P}\left\{|\mathcal{F}(\mathbf{z}) - \mathbb{E}\mathcal{F}(\mathbf{z})| > t\right\} \leq Ce^{-c\,t^q}$$

Why does it work on real data?

- Real data are **unlikely close** to Gaussian.
- However, using **GANs**, it is possible to

$$\text{Real data} \approx \underbrace{\mathcal{F}_1 \circ \mathcal{F}_2 \cdots \circ \mathcal{F}_N}_{\textit{Lipschitz}\text{ operations}}(\text{Gaussian})$$

- On the other hand, a **concentrated** vector $\mathbf{z}$ is defined as
For any *Lipschitz* observation $\mathcal{F} : \mathbb{R}^p \to \mathbb{R}$,

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

Why does it work on real data?

- Real data are **unlikely close** to Gaussian.
- However, using **GANs**, it is possible to

$$\text{Real data} \approx \underbrace{\mathcal{F}_1 \circ \mathcal{F}_2 \cdots \circ \mathcal{F}_N}_{\textit{Lipschitz} \text{ operations}}(\text{Gaussian})$$

- On the other hand, a **concentrated** vector $\mathbf{z}$ is defined as
For any *Lipschitz* observation $\mathcal{F} : \mathbb{R}^p \to \mathbb{R}$,

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

- Example: $\mathcal{G}$(Gaussian) with $\mathcal{G} : \mathbb{R}^p \to \mathbb{R}^d$ Lipschitz.
[Ledoux'05]

Why does it work on real data?

- Real data are **unlikely close** to Gaussian.
- However, using **GANs**, it is possible to

$$\text{Real data} \approx \underbrace{\mathcal{F}_1 \circ \mathcal{F}_2 \cdots \circ \mathcal{F}_N}_{\textit{Lipschitz}\text{ operations}}(\text{Gaussian})$$

- On the other hand, a **concentrated** vector $\mathbf{z}$ is defined as
  For any *Lipschitz* observation $\mathcal{F} : \mathbb{R}^p \to \mathbb{R}$,

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

- Example: $\mathcal{G}$(Gaussian) with $\mathcal{G} : \mathbb{R}^p \to \mathbb{R}^d$ Lipschitz. [Ledoux'05]

$\Rightarrow$ Consider data as **concentrated** vectors.

Outline

Concentrated Mixture Model

Consider data distributed in k classes as

$$\mathbf{X} = [\underbrace{\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}}_{\in \mathcal{O}(e^{-\cdot^{q_1}})}, \underbrace{\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_{n_2}}_{\in \mathcal{O}(e^{-\cdot^{q_2}})}, \ldots, \underbrace{\mathbf{x}_{n-n_k+1}, \ldots, \mathbf{x}_n}_{\in \mathcal{O}(e^{-\cdot^{q_k}})}]$$

Concentrated Mixture Model

Consider data distributed in k classes as

$$\mathbf{X} = [\underbrace{\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}}_{\in \mathcal{O}(e^{-\cdot^{q_1}})}, \underbrace{\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_{n_2}}_{\in \mathcal{O}(e^{-\cdot^{q_2}})}, \ldots, \underbrace{\mathbf{x}_{n-n_k+1}, \ldots, \mathbf{x}_n}_{\in \mathcal{O}(e^{-\cdot^{q_k}})}]$$

Recall

$$\mathbf{m} = \sum_{\ell=1}^{k} \frac{n_\ell}{n} \mathbf{m}_\ell, \; \bar{\mathbf{m}}_\ell = \mathbf{m}_\ell - \mathbf{m}, \;\; \mathbf{C} = \sum_{\ell=1}^{k} \frac{n_\ell}{n} \mathbf{C}_\ell, \; \bar{\mathbf{C}}_\ell = \mathbf{C}_\ell - \mathbf{C}$$

Assumption (Growth rate)

As $n \to \infty$,

1. **Data:** $\frac{p}{n} \to c_0 \in (0, \infty)$, $\frac{n_\ell}{n} \to c_\ell \in (0, 1)$
2. **Mean:** $\|\bar{\mathbf{m}}_\ell\| = \mathcal{O}(1)$, $\mathbb{E}\|\mathbf{x}_i\| = \mathcal{O}(\sqrt{p})$
3. **Covariance:** $\|\bar{\mathbf{C}}_\ell\| = \mathcal{O}(1)$, $\operatorname{tr}\bar{\mathbf{C}}_\ell = \mathcal{O}(\sqrt{p})$, $\operatorname{tr}\bar{\mathbf{C}}_a\bar{\mathbf{C}}_b = \mathcal{O}(p)$

MP Law Still Valid for Concentrated Vectors!

Theorem (Spectrum of $\frac{1}{p}\mathbf{Z}^\intercal \mathbf{Z}$ [Louart'18])

Let $z \in \mathbb{C}^+$ and define the resolvent matrix

$$\mathbf{Q}_\delta(z) \equiv \left(\sum_{\ell=1}^{k} c_\ell \frac{\mathbf{C}_\ell}{1+\delta_\ell(z)} - z\mathbf{I}_p \right)^{-1}$$

MP Law Still Valid for Concentrated Vectors!

Theorem (Spectrum of $\frac{1}{p}\mathbf{Z}^\intercal \mathbf{Z}$ [Louart'18])

Let $z \in \mathbb{C}^+$ and define the resolvent matrix

$$\mathbf{Q}_\delta(z) \equiv \left(\sum_{\ell=1}^{k} c_\ell \frac{\mathbf{C}_\ell}{1 + \delta_\ell(z)} - z\mathbf{I}_p \right)^{-1}$$

where $\delta_\ell(z)$ is the unique solution of the fixed point equation

$$\delta_\ell(z) = \frac{1}{n}(\mathbf{C}_\ell \mathbf{Q}_\delta(z)).$$

MP Law Still Valid for Concentrated Vectors!

Theorem (Spectrum of $\frac{1}{p}\mathbf{Z}^\intercal\mathbf{Z}$ [Louart'18])

Let $z \in \mathbb{C}^+$ and define the resolvent matrix

$$\mathbf{Q}_\delta(z) \equiv \left(\sum_{\ell=1}^{k} c_\ell \frac{\mathbf{C}_\ell}{1+\delta_\ell(z)} - z\mathbf{I}_p \right)^{-1}$$

where $\delta_\ell(z)$ is the unique solution of the fixed point equation

$$\delta_\ell(z) = \frac{1}{n}(\mathbf{C}_\ell \mathbf{Q}_\delta(z)).$$

Then the e.s.d. $\nu_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{\lambda_i}$ a.s. converges to a probability measure ν defined on a compact support $\mathcal{S}$ and having density

$$f(x) = \lim_{\epsilon\to 0,\, p\to\infty} \frac{1}{p} \mathcal{I}m\, Tr \mathbf{Q}_\delta(x+i\epsilon)$$

Random Matrix Equivalent

- **We still have:**

$$\boxed{\max_{1\leq i\neq j\leq n}\left\{\left|\frac{1}{p}\|\mathbf{x}_i-\mathbf{x}_j\|^2-\tau\right|\right\}\xrightarrow{a.s.}0}$$

where $\tau = \frac{2}{p}\operatorname{tr}\mathbf{C}$.

Random Matrix Equivalent

- **We still have:**

$$\max_{1\leq i\neq j\leq n}\left\{\left|\frac{1}{p}\|\mathbf{x}_i-\mathbf{x}_j\|^2-\tau\right|\right\}\xrightarrow{a.s.}0$$

where $\tau=\frac{2}{p}\mathrm{tr}\mathbf{C}$.

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

$$\mathbf{K}\propto\underbrace{\mathbf{J}\mathbf{A}\mathbf{J}^\mathsf{T}}_{\textit{Information}}+\underbrace{f'(\tau)\mathbf{Z}^\mathsf{T}\mathbf{Z}+*}_{\textit{Noise}}$$

Random Matrix Equivalent

- **We still have:**

$$\max_{1\leq i\neq j\leq n}\left\{\left|\frac{1}{p}\|\mathbf{x}_i-\mathbf{x}_j\|^2-\tau\right|\right\}\xrightarrow{a.s.}0$$

where $\tau=\frac{2}{p}\mathrm{tr}\mathbf{C}$.

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

$$\mathbf{K}\propto\underbrace{\mathbf{JAJ}^\mathsf{T}}_{\textit{Information}}+\underbrace{f'(\tau)\mathbf{Z}^\mathsf{T}\mathbf{Z}+*}_{\textit{Noise}}+\underbrace{\Delta(\text{Gaussian},\text{Concentrated})}_{\text{Non-informative 4th order moments}}$$

Random Matrix Equivalent

- **We still have:**

$$\max_{1\le i\ne j\le n}\left\{\left|\frac{1}{p}\|\mathbf{x}_i-\mathbf{x}_j\|^2-\tau\right|\right\}\xrightarrow{a.s.}0$$

where $\tau=\frac{2}{p}\mathrm{tr}\mathbf{C}$.

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

$$\mathbf{K}\propto\underbrace{\mathbf{J}\mathbf{A}\mathbf{J}^\mathsf{T}}_{Information}+\underbrace{f'(\tau)\mathbf{Z}^\mathsf{T}\mathbf{Z}+*}_{Noise}+\underbrace{\Delta(\text{Gaussian},\text{Concentrated})}_{\text{Non-informative 4th order moments}}$$

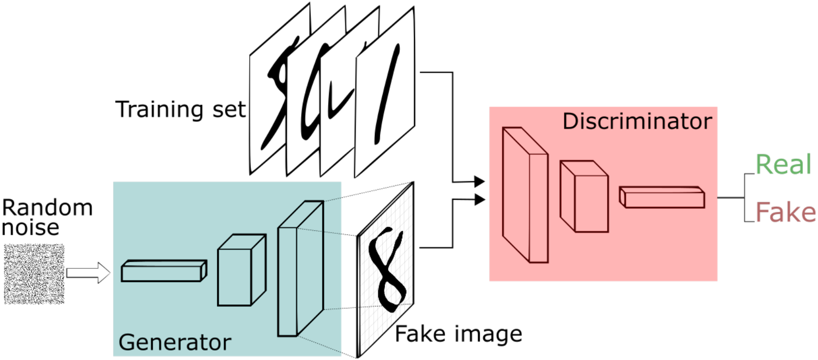
where $\mathbf{A}\propto f'(\tau)\mathbf{M}^\mathsf{T}\mathbf{M}+f''(\tau)[\mathbf{t}\mathbf{t}^\mathsf{T}+\mathbf{T}]$, and

$$\mathbf{J}=[\mathbf{j}_1,\ldots,\mathbf{j}_k],\;\mathbf{M}=[\bar{\mathbf{m}}_1,\ldots,\bar{\mathbf{m}}_k]$$

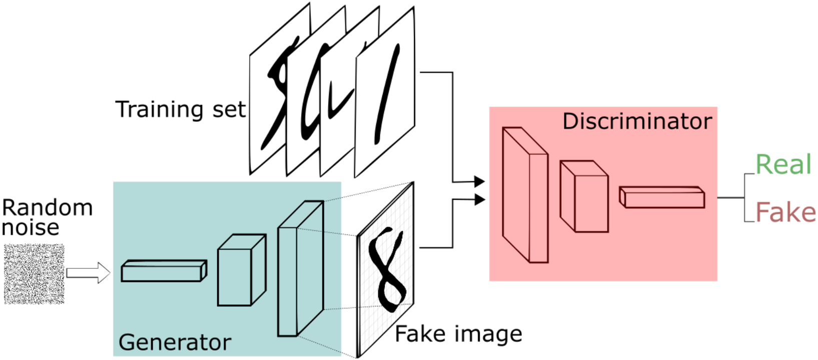
$$\mathbf{t}=\left\{\frac{\mathrm{tr}\bar{\mathbf{C}}_\ell}{\sqrt{p}}\right\}_{\ell=1}^k,\;\mathbf{T}=\left\{\frac{\mathrm{tr}\bar{\mathbf{C}}_a\bar{\mathbf{C}}_b}{p}\right\}_{a,b=1}^k$$

Outline

Application to GAN-generated Images

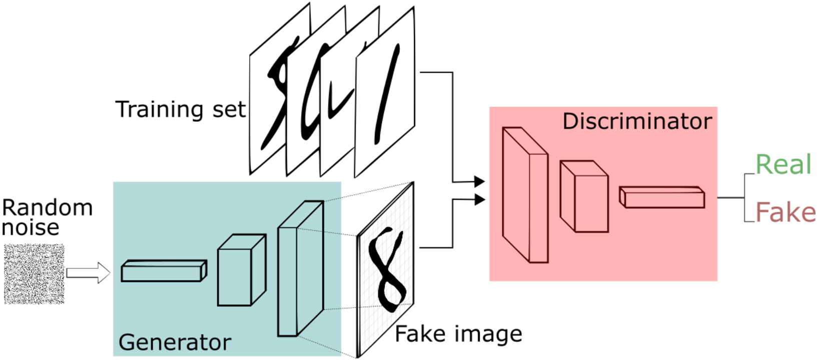


Application to GAN-generated Images



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

Application to GAN-generated Images



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

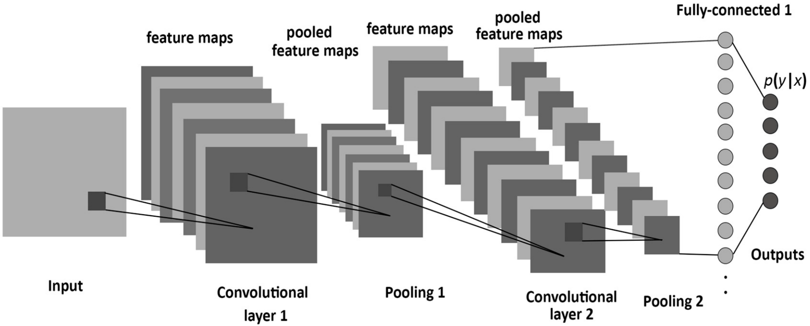
We generate data as

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

Application to GAN-generated Images

Figure 7: Images generated by a GAN model **[Brock et al.'18]**.

CNN Representations



We consider data as (commonly used in Computer Vision)

$$\mathbf{x}_i = \underbrace{\mathcal{R} \circ \mathcal{G}}_{\text{Lipschitz}}(\text{Gaussian})$$

Application to GAN-generated Images

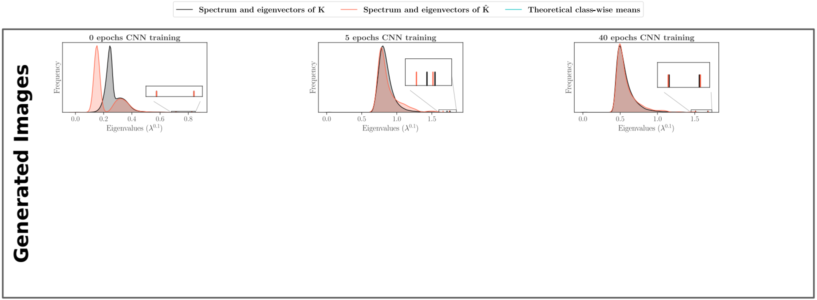


Figure 8: Spectral clustering on GAN-generated images.
$f(t) = \exp(-t)$, $k = 3$, $n = 3000$ and $p = 1024$.

Application to GAN-generated Images

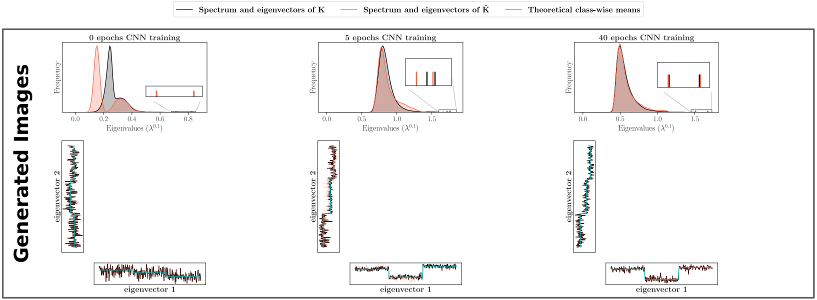


Figure 9: Spectral clustering on GAN-generated images.
$f(t) = \exp(-t)$, $k = 3$, $n = 3000$ and $p = 1024$.

Application to GAN-generated Images

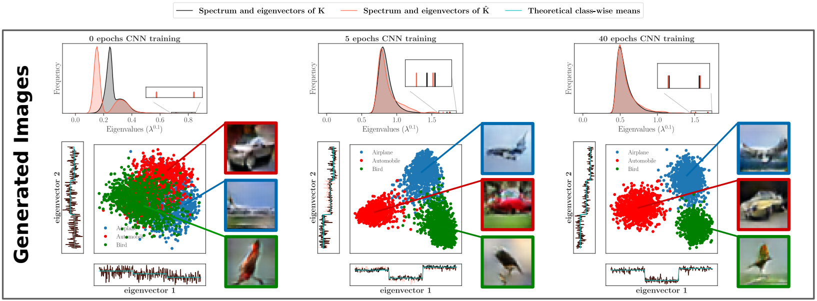


Figure 10: Spectral clustering on GAN-generated images.
$f(t) = \exp(-t)$, $k = 3$, $n = 3000$ and $p = 1024$.

Application to Real Images

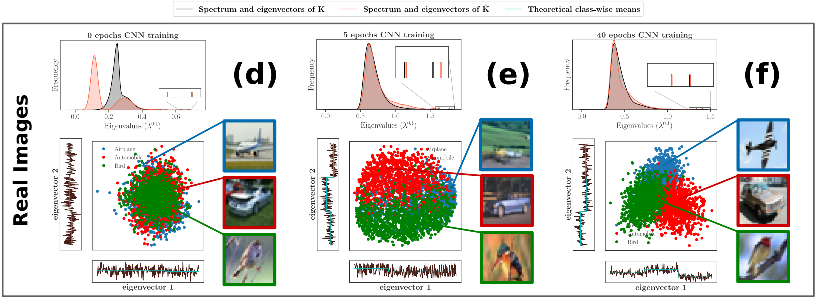


Figure 11: Spectral clustering on real images. $f(t) = \exp(-t)$, $k = 3$, $n = 3000$ and $p = 1024$.

Summary & Perspectives

- RMT allows for the theoretical understanding of ML methods for *concentrated vectors*.

Summary & Perspectives

- RMT allows for the theoretical understanding of ML methods for *concentrated vectors*.
- Real data behave similar to concentrated vectors.

Summary & Perspectives

- RMT allows for the theoretical understanding of ML methods for *concentrated vectors*.
- Real data behave similar to concentrated vectors.

$\Rightarrow$ RMT is relevant for ML!

Summary & Perspectives

- RMT allows for the theoretical understanding of ML methods for *concentrated vectors*.
- Real data behave similar to concentrated vectors.

$\Rightarrow$ RMT is relevant for ML!

- Generalize to other ML tasks (Classification, Regression etc.).

Summary & Perspectives

- RMT allows for the theoretical understanding of ML methods for *concentrated vectors*.
- Real data behave similar to concentrated vectors.

$\Rightarrow$ RMT is relevant for ML!

- Generalize to other ML tasks (Classification, Regression etc.).
- Use the concentration framework to understand the dynamics of neural networks and GANs.

Thanks for your attention!