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

Introduction

Behavior of the Gram Matrix for Gaussian Vectors

Notion of Concentrated Vectors
  Definition and Basic Properties
  GAN Data : An Example of Concentrated Vectors

Behavior of the Gram Matrix for Concentrated Vectors

Behavior of Kernel Matrices for Concentrated Vectors

Application to CNN Representations of GAN Images
Outline

Introduction

Behavior of the Gram Matrix for Gaussian Vectors

Notion of Concentrated Vectors
  Definition and Basic Properties
  GAN Data : An Example of Concentrated Vectors

Behavior of the Gram Matrix for Concentrated Vectors

Behavior of Kernel Matrices for Concentrated Vectors

Application to CNN Representations of GAN Images
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 \( x_i^\top x_j \)

Quite naturally, the Gram matrix \( X^\top X \) 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

Introduction

Behavior of the Gram Matrix for Gaussian Vectors

Notion of Concentrated Vectors
  Definition and Basic Properties
  GAN Data: An Example of Concentrated Vectors

Behavior of the Gram Matrix for Concentrated Vectors

Behavior of Kernel Matrices for Concentrated Vectors

Application to CNN Representations of GAN Images
Let us assume \( x_i \sim \mathcal{N}(0, I_p) \)

**Figure:** Eigenvalues distribution of \( \frac{1}{p} X^T X \) for \( n = p = 1000 \).
Definition (Empirical Spectral Density)

The empirical spectral density (e.s.d.) $\mu_n$ 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. $\mu_n$ of $\frac{1}{p}X^T X$ satisfies

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

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

$$
f_c(x) = \frac{1}{2\pi cx} \sqrt{(x - \lambda^-)(\lambda^+ - x)}
$$
Gaussian Mixture (Spiked Model)

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

$$X = \begin{bmatrix} x_1, \ldots, x_{\frac{n}{2}}, x_{\frac{n}{2} + 1}, \ldots, x_n \end{bmatrix} \sim \mathcal{N}(+\mu, I_p) \quad \text{and} \quad \begin{bmatrix} x_{\frac{n}{2}}, \ldots, x_n \end{bmatrix} \sim \mathcal{N}(-\mu, I_p)$$

- We can write

$$X = \mu y^T + Z$$

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

$$\frac{1}{p} X^T X = \underbrace{\|\mu\|^2 \bar{y} \bar{y}^T}_{\text{Information (low-rank)}} + \frac{1}{p} Z^T Z + \ast \quad \text{where } \bar{y} = y/\sqrt{p}$$
Gaussian Mixture (Spiked Model)

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

Visible if $\|m\|^2 \geq \sqrt{c}$
Gaussian Mixture (Spiked Model)

Figure: Eigenvalues distribution of $\frac{1}{p} X^T X$ along with its dominant eigenvector for $n = p = 1000$. Visible if $\|m\|^2 \geq \sqrt{c}$.
Some RMT Results on Spiked Models

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

Let

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

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

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

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

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

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

In particular,

$$
|\hat{y}^\top y|^2 \xrightarrow{a.s.} \frac{1 - c\|\mu\|^{-4}}{1 + c\|\mu\|^{-2}} \cdot \mathbf{1}\|\mu\|^2 > \sqrt{c}.
$$
Some RMT Results on Spiked Models

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

Introduction

Behavior of the Gram Matrix for Gaussian Vectors

**Notion of Concentrated Vectors**
  - Definition and Basic Properties
  - GAN Data: An Example of Concentrated Vectors

Behavior of the Gram Matrix for Concentrated Vectors

Behavior of Kernel Matrices for Concentrated Vectors

Application to CNN Representations of GAN Images
Outline

Introduction

Behavior of the Gram Matrix for Gaussian Vectors

Notion of Concentrated Vectors
  Definition and Basic Properties
    GAN Data: An Example of Concentrated Vectors

Behavior of the Gram Matrix for Concentrated Vectors

Behavior of Kernel Matrices for Concentrated Vectors

Application to CNN Representations of GAN Images
Notion of Concentrated Vectors

▶ **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 \(z \in E\) is \(q\)-exponentially **concentrated** if for any 1-Lipschitz\(^1\) function \(F : \mathbb{R}^p \to \mathbb{R}\), there exists \(C, c > 0\) such that

\[
\mathbb{P}\{\|F(z) - \mathbb{E}F(z)\| > t\} \leq Ce^{-c t^q} \quad \text{denoted} \quad z \in O(e^{-\cdot q})
\]

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

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

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

“Concentrated vectors are stable through Lipschitz maps.”

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

Introduction

Behavior of the Gram Matrix for Gaussian Vectors

Notion of Concentrated Vectors
  Definition and Basic Properties
  GAN Data: An Example of Concentrated Vectors

Behavior of the Gram Matrix for Concentrated Vectors

Behavior of Kernel Matrices for Concentrated Vectors

Application to CNN Representations of GAN Images
GAN Data: An Example of Concentrated Vectors

\begin{align*}
\min_G \max_D & \mathbb{E}_{x \sim p(x)} [\log D(x)] + \mathbb{E}_{z \sim p(z)} [\log (1 - D(G(z)))] \\
\end{align*}

Once the Generator is trained, we generate data as

**Generated Image** = $G(\text{Gaussian})$
GAN Data: An Example of Concentrated Vectors

\[ \text{GAN Data} = F_1 \circ F_2 \circ \cdots \circ F_N(\text{Gaussian}) \]

where the $F_i$’s correspond to Fully Connected layers, Convolutional layers, Pooling and activation functions, residual connections or Batch Normalisation.

⇒ The $F_i$’s are essentially \textit{Lipschitz} operations.
GAN Data: An Example of Concentrated Vectors

- **Fully Connected Layers and Convolutional Layers** are affine operations:
  \[ \mathcal{F}_i(x) = W_i x + b_i, \]
  \[ \text{and } \|\mathcal{F}_i\|_{lip} = \sup_{u \neq 0} \frac{\|W_i u\|_p}{\|u\|_p}, \text{ for any } p\text{-norm.} \]

- **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 \( C_1, C_2, \ldots, C_k \) as

\[
X = \begin{bmatrix}
  x_1, \ldots, x_{n_1}, & x_{n_1+1}, \ldots, x_{n_2}, & \cdots, & x_{n-n_k+1}, \ldots, x_n
\end{bmatrix} \in \mathbb{R}^{p \times n}
\]

\[
\in \mathcal{O}(e^{-\cdot q_1}) \quad \in \mathcal{O}(e^{-\cdot q_2}) \quad \in \mathcal{O}(e^{-\cdot q_k})
\]

Denote

\[
\mu_\ell = \mathbb{E}_{x_i \in C_\ell} [x_i], \quad C_\ell = \mathbb{E}_{x_i \in C_\ell} [x_i x_i^T]
\]

**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

Introduction

Behavior of the Gram Matrix for Gaussian Vectors

Notion of Concentrated Vectors
  Definition and Basic Properties
  GAN Data: An Example of Concentrated Vectors

Behavior of the Gram Matrix for Concentrated Vectors

Behavior of Kernel Matrices for Concentrated Vectors

Application to CNN Representations of GAN Images
Behavior of the Gram Matrix for Concentrated Vectors

Let

\[ G = \frac{1}{p} X^T X = \frac{1}{p} J M^T M J^T + \frac{1}{p} Z^T Z + * + o_p(1) \]

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

\[ L = \frac{1}{n} \sum_i \delta_{\lambda_i}, \ m_L(z) = \int_{\lambda} \frac{dL(\lambda)}{\lambda - z} = \frac{1}{n} \text{tr} (Q(-z)) \]

\[ U U^T = \frac{1}{2\pi i} \oint_{\gamma} Q(-z) dz \]

⇒ Analyse the behavior of the resolvent \( Q(z) = (G + z I_n)^{-1} \).
Behavior of the Gram Matrix for Concentrated Vectors

**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,

$$\| \mathbb{E}[Q(z)] - \tilde{Q}(z) \| = \mathcal{O} \left( \sqrt{\frac{\log p}{p}} \right)$$

where $\tilde{Q}(z) = \frac{1}{z} \Lambda(z) + \frac{1}{pz} J \Omega(z) J^T$

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

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

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

$$\delta_{\ell}(z) = \text{tr} \left( C_{\ell} \left( \frac{1}{k} \sum_{j=1}^k \frac{C_{j}}{1 + \delta_{j}(z) + zl_p} \right)^{-1} \right) \text{ for each } \ell \in [k].$$
Behavior of the Gram Matrix for Concentrated Vectors

Theorem

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

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

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

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

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

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

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

Introduction

Behavior of the Gram Matrix for Gaussian Vectors

Notion of Concentrated Vectors
  Definition and Basic Properties
  GAN Data : An Example of Concentrated Vectors

Behavior of the Gram Matrix for Concentrated Vectors

Behavior of Kernel Matrices for Concentrated Vectors

Application to CNN Representations of GAN Images
Problem Statement

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

\[
K = \left\{ f \left( \frac{1}{p} \| x_i - x_j \|_2^2 \right) \right\}_{i,j=1}^n
\]

Intuition (from small dimensions)

$K$ mainly low rank with class information in eigenvectors.
Small Dimension vs High Dimension!

\[ Z = \begin{pmatrix} \gg 1 & \ll 1 & \ll 1 \\ \ll 1 & \gg 1 & \ll 1 \\ \ll 1 & \ll 1 & \gg 1 \end{pmatrix} \]
Key Observation: The between and within class vectors are “equidistant” in high-dimension.

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

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

Taylor Expanding \( K \) entry-wise leads to

\[
K \propto \begin{cases} \text{Information} \ \ JAJ^T \end{cases} + f'(\tau)Z^TZ + * \begin{cases} \text{Noise} \end{cases}
\]

where \( A \propto f'(\tau)M^TM + f''(\tau) [tt^T + T] \), and

\[
J = [j_1, \ldots, j_k], \quad M = [\bar{m}_1, \ldots, \bar{m}_k], \quad t = \left\{ \frac{\text{tr} \bar{C}_{\ell}}{\sqrt{p}} \right\}_{\ell=1}^k, \quad T = \left\{ \frac{\text{tr} \bar{C}_a \bar{C}_b}{p} \right\}_{a,b=1}^k
\]

Result: Only first and second order statistics matter!
Outline

Introduction

Behavior of the Gram Matrix for Gaussian Vectors

Notion of Concentrated Vectors
  Definition and Basic Properties
  GAN Data: An Example of Concentrated Vectors

Behavior of the Gram Matrix for Concentrated Vectors

Behavior of Kernel Matrices for Concentrated Vectors

Application to CNN Representations of GAN Images
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.
Application to CNN Representations of GAN Images

\[ z \sim \mathcal{N}(0, I) \]

Generator

Discriminator

Real / Fake

Lipschitz operation

Representation Network

Concentrated Vectors
Application to CNN Representations of GAN Images

GAN Images

Real Images
Application to CNN Representations of GAN Images

**GAN Images**

resnet50 ($p = 2048$)

Real Images

resnet50 ($p = 2048$)

vgg16 ($p = 4096$)

densenet201 ($p = 1920$)

densenet201 ($p = 1920$)
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/