Pourquoi les matrices aléatoires expliquent l'apprentissage ? Un argument d'universalité offert par les GANs

(Gretsi'2019, Lille)

Mohamed El Amine SEDDIK, Mohamed TAMAAZOUSTI, Romain COUILLET

CEA List, France
CentraleSupélec, L2S, Université ParisSaclay, France
GSTATS IDEX DataScience Chair, GIPSA-lab, Université Grenoble–Alpes, France.

29 août 2019

Outline

Introduction

Comportement de la matrice de Gram de vecteurs Gaussiens

Notion de vecteurs concentrés

- Definition et propriétés
- Données GAN : Un exemple de vecteurs concentrés

Comportement de la matrice de Gram de vecteurs concentrés

Application aux representations CNN des images GAN

Outline

Introduction

Comportement de la matrice de Gram de vecteurs Gaussiens

Notion de vecteurs concentrés
- Definition et propriétés
- Données GAN : Un exemple de vecteurs concentrés

Comportement de la matrice de Gram de vecteurs concentrés

Application aux representations CNN des images GAN

En apprentissage machine (ML),

En apprentissage machine (ML),

- Etant donné un ensemble de données

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

En apprentissage machine (ML),

- Etant donné un ensemble de données

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

- On aimerait réaliser des tâches de

Régression, Classification, Regroupement etc.

Introduction

En apprentissage machine (ML),

- Etant donné un ensemble de données

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

- On aimerait réaliser des tâches de

Régression, Classification, Regroupement etc.

- Et au cœur de ces tâches, on calcule des **similarités**

Introduction

En apprentissage machine (ML),

- Etant donné un ensemble de données

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

- On aimerait réaliser des tâches de

Régression, Classification, Regroupement etc.

- Et au cœur de ces tâches, on calcule des **similarités**

Par exemple : le produit scalaire $x_i^\intercal x_j$

Introduction

En apprentissage machine (ML),

- Etant donné un ensemble de données

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

- On aimerait réaliser des tâches de

Régression, Classification, Regroupement etc.

- Et au cœur de ces tâches, on calcule des **similarités**

Par exemple : le produit scalaire $x_i^\intercal x_j$

Naturellement, la matrice de Gram $X^\intercal X$ apparaît en ML.

En apprentissage machine (ML),

- Etant donné un ensemble de données

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

- On aimerait réaliser des tâches de

Régression, Classification, Regroupement etc.

- Et au cœur de ces tâches, on calcule des **similarités**

Par exemple : le produit scalaire $x_i^\intercal x_j$

Naturellement, la matrice de Gram $X^\intercal X$ apparaît en ML.

- **Comment se comporte t-elle ?**

Outline

Comportement de la matrice de Gram de vecteurs Gaussiens

- Si on suppose tous les $x_i \sim \mathcal{N}(0, \boldsymbol{I}_p)$

Comportement de la matrice de Gram de vecteurs Gaussiens

- Si on suppose tous les $x_i \sim \mathcal{N}(0, I_p)$

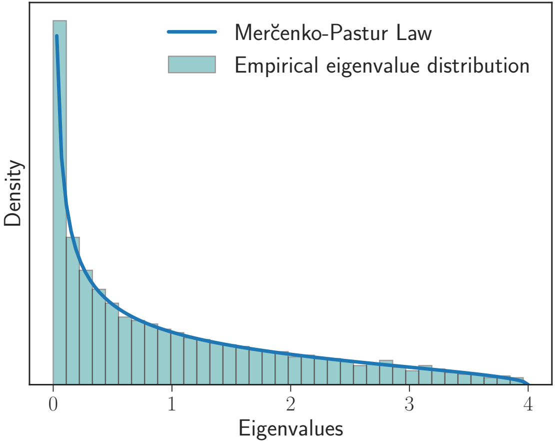


Figure: Histogramme des valeurs propres de $\frac{1}{p} X^{\mathsf{T}} X$ pour $n = p = 1000$.

Définition (Densité Spectrale Empirique)

La densité spectrale empirique (e.s.d.) μ_n *d'une matrice hermitienne* $A_n \in \mathbb{R}^{n\times n}$ *est donnée par* $\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(A_n)}$.

Loi de Marčenko–Pastur **[Marčenko,Pastur'67]**

Définition (Densité Spectrale Empirique)

La densité spectrale empirique (e.s.d.) μ_n *d'une matrice hermitienne* $A_n \in \mathbb{R}^{n\times n}$ *est donnée par* $\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(A_n)}$.

Théorème (Loi de Marčenko–Pastur)

Soit $X \in \mathbb{R}^{p\times n}$ *avec des entrées i.i.d. de moyenne nulle, et de variance* 1.
Quand $p, n \to \infty$ *avec* $n/p \to c \in (0, \infty)$*, la e.s.d.* μ_n *de* $\frac{1}{p}X^\intercal X$ *satisfait*

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

où μ_c *est une mesure détirministe de densité continue* f_c *sur un support compacte* $[\lambda^-, \lambda^+] = [(1-\sqrt{c})^2, (1+\sqrt{c})^2]$

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

Mélange Gaussien (Modèle spike)

- Soit $\mu \in \mathbb{R}^p$ tel que $\|\mu\| = \mathcal{O}(1)$

Mélange Gaussien (Modèle spike)

- Soit $\mu \in \mathbb{R}^p$ tel que $\|\mu\| = \mathcal{O}(1)$
- Considérons

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

Mélange Gaussien (Modèle spike)

- Soit $\mu \in \mathbb{R}^p$ tel que $\|\mu\| = \mathcal{O}(1)$
- Considérons

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

- Qu'on peut écrire

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

où $y \in \{+1, -1\}^n$ est le vecteur de labèles et Z a des entrées i.i.d. $\mathcal{N}(0, 1)$.

Mélange Gaussien (Modèle spike)

- Soit $\mu \in \mathbb{R}^p$ tel que $\|\mu\| = \mathcal{O}(1)$
- Considérons

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

- Qu'on peut écrire

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

où $y \in \{+1, -1\}^n$ est le vecteur de labèles et Z a des entrées i.i.d. $\mathcal{N}(0,1)$.
- On a donc

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

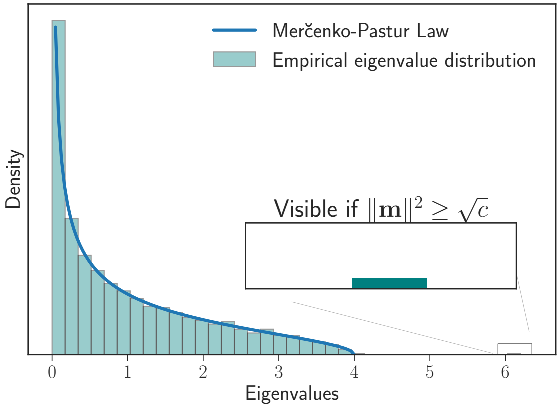


Figure: Histogramme des valeurs propres de $\frac{1}{p}X^{\intercal}X$ pour $n = p = 1000$.

Mélange Gaussien (Modèle spike)

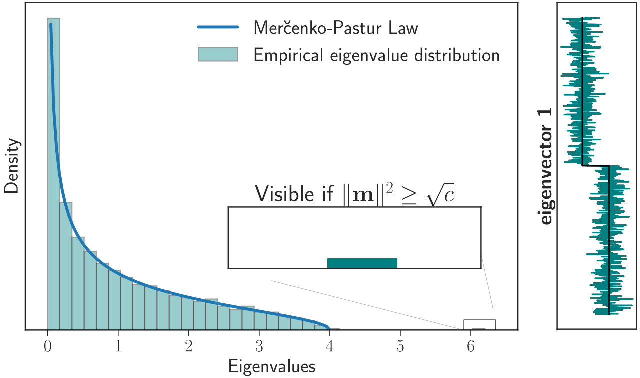


Figure: Histogramme des valeurs propres de $\frac{1}{p}X^{\mathsf{T}}X$ et son vecteur propre dominant pour $n = p = 1000$.

Théorème (**[Baik, Silverstein'06], [Paul'07]**)

Let

- *Soit Z avec des entées i.i.d. de moyenne nulle, variance 1 et $\mathbb{E}|Z_{ij}|^4 < \infty$*
- $X = my^\intercal + Z$

Théorème (**[Baik, Silverstein'06], [Paul'07]**)

Let

- *Soit* Z *avec des entées i.i.d. de moyenne nulle, variance* 1 *et* $\mathbb{E}|Z_{ij}|^4 < \infty$
- $X = my^\mathsf{T} + Z$

Alors, quand $p, n \to$ *avec* $n/p \to c$*,*

Théorème (**[Baik, Silverstein'06], [Paul'07]**)

Let

- *Soit Z avec des entées i.i.d. de moyenne nulle, variance* 1 *et* $\mathbb{E}|Z_{ij}|^4 < \infty$
- $X = my^\mathsf{T} + Z$

Alors, quand $p, n \to$ avec $n/p \to c$,

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

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

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

Let

- *Soit Z avec des entées i.i.d. de moyenne nulle, variance 1 et $\mathbb{E}|Z_{ij}|^4 < \infty$*
- $X = my^\mathsf{T} + Z$

Alors, quand $p, n \to$ avec $n/p \to c$,

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

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

- *Pour $a, b \in \mathbb{R}^p$ détirministes et $\hat{y}$ le vecteur propre correspondant à $\lambda_{\max}\left(\frac{1}{p}X^\mathsf{T}X\right)$,*

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

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

Let

- *Soit* Z *avec des entées i.i.d. de moyenne nulle, variance* 1 *et* $\mathbb{E}|Z_{ij}|^4 < \infty$
- $X = my^\mathsf{T} + Z$

Alors, quand $p, n \to$ *avec* $n/p \to c$*,*

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

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

- *Pour* $a, b \in \mathbb{R}^p$ *détirministes et* $\hat{y}$ *le vecteur propre correspondant à* $\lambda_{\max}\left(\frac{1}{p}X^\mathsf{T}X\right)$*,*

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

En particulier,

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

Quelques résultats sur les modèles spike

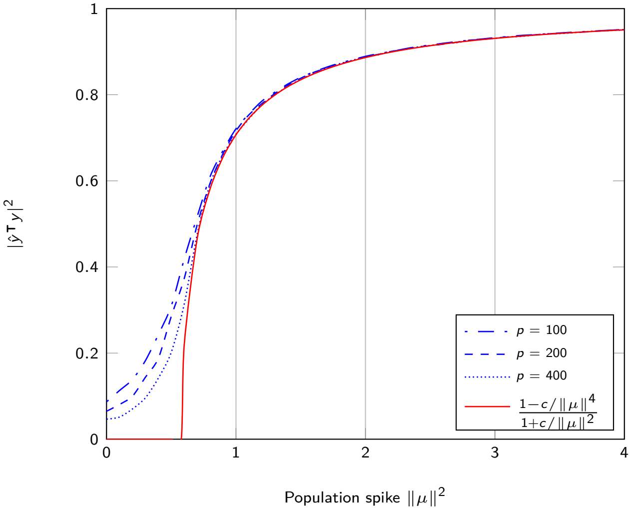


Figure: $|\hat{y}^\mathsf{T} y|^2$ simulé et valeur limite, $p/n = 1/3$, en variant $\|\mu\|^2$.

Outline

Outline

Notion de vecteurs concentrés

- **Observation**: La TMA prédit les performances en grande dimension sous des hypothèses Gaussiennes sur les données.

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

Notion de vecteurs concentrés

- **Observation**: La TMA prédit les performances en grande dimension sous des hypothèses Gaussiennes sur les données.
- **MAIS** les données réelles sont difficilement **assimilable** à des vecteurs Gaussiens.

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

Notion de vecteurs concentrés

- **Observation**: La TMA prédit les performances en grande dimension sous des hypothèses Gaussiennes sur les données.
- **MAIS** les données réelles sont difficilement **assimilable** à des vecteurs Gaussiens.
- Il se trouve que les vecteurs Gaussiens appartiennent à une classe plus large et plus utile de vecteurs aléatoires.

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

- **Observation**: La TMA prédit les performances en grande dimension sous des hypothèses Gaussiennes sur les données.
- **MAIS** les données réelles sont difficilement **assimilable** à des vecteurs Gaussiens.
- Il se trouve que les vecteurs Gaussiens appartiennent à une classe plus large et plus utile de vecteurs aléatoires.

Definition

Etant donné un espace normé $(E, \|\cdot\|_E)$ et $q \in \mathbb{R}$, un vecteur aléatoire $\mathbf{z} \in E$ est q-exponentiellement **concentré** si pour toute fonction 1-**Lipschitz**[1] $\mathcal{F} : \mathbb{R}^p \to \mathbb{R}$, il existe $C, c > 0$ tels que

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

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

Notion de vecteurs concentrés

- **Observation**: La TMA prédit les performances en grande dimension sous des hypothèses Gaussiennes sur les données.
- **MAIS** les données réelles sont difficilement **assimilable** à des vecteurs Gaussiens.
- Il se trouve que les vecteurs Gaussiens appartiennent à une classe plus large et plus utile de vecteurs aléatoires.

Definition

Etant donné un espace normé $(E, \|\cdot\|_E)$ et $q \in \mathbb{R}$, un vecteur aléatoire $\mathbf{z} \in E$ est q-exponentiellement **concentré** si pour toute fonction 1-**Lipschitz**[1] $\mathcal{F} : \mathbb{R}^p \to \mathbb{R}$, il existe $C, c > 0$ tels que

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

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

Notion de vecteurs concentrés

- **Observation**: La TMA prédit les performances en grande dimension sous des hypothèses Gaussiennes sur les données.
- **MAIS** les données réelles sont difficilement **assimilable** à des vecteurs Gaussiens.
- Il se trouve que les vecteurs Gaussiens appartiennent à une classe plus large et plus utile de vecteurs aléatoires.

Definition

Etant donné un espace normé $(E, \|\cdot\|_E)$ et $q \in \mathbb{R}$, un vecteur aléatoire $\mathbf{z} \in E$ est q-exponentiellement **concentré** si pour toute fonction 1-**Lipschitz**[1] $\mathcal{F} : \mathbb{R}^p \to \mathbb{R}$, il existe $C, c > 0$ tels que

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

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

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

Notion de vecteurs concentrés

- **Observation**: La TMA prédit les performances en grande dimension sous des hypothèses Gaussiennes sur les données.
- **MAIS** les données réelles sont difficilement **assimilable** à des vecteurs Gaussiens.
- Il se trouve que les vecteurs Gaussiens appartiennent à une classe plus large et plus utile de vecteurs aléatoires.

Definition

Etant donné un espace normé $(E, \|\cdot\|_E)$ et $q \in \mathbb{R}$, un vecteur aléatoire $\mathbf{z} \in E$ est q-exponentiellement **concentré** si pour toute fonction 1-**Lipschitz**[1] $\mathcal{F} : \mathbb{R}^p \to \mathbb{R}$, il existe $C, c > 0$ tels que

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

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

(P2) Si $X \in \mathcal{O}(e^{-\cdot^q})$ et $\mathcal{G}$ est $\|\mathcal{G}\|_{lip}$-**Lipschitz**, alors

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

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

Notion de vecteurs concentrés

- **Observation**: La TMA prédit les performances en grande dimension sous des hypothèses Gaussiennes sur les données.
- **MAIS** les données réelles sont difficilement **assimilable** à des vecteurs Gaussiens.
- Il se trouve que les vecteurs Gaussiens appartiennent à une classe plus large et plus utile de vecteurs aléatoires.

Definition

Etant donné un espace normé $(E, \|\cdot\|_E)$ et $q \in \mathbb{R}$, un vecteur aléatoire $\mathbf{z} \in E$ est q-exponentiellement **concentré** si pour toute fonction 1-**Lipschitz**[1] $\mathcal{F} : \mathbb{R}^p \to \mathbb{R}$, il existe $C, c > 0$ tels que

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

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

(P2) Si $X \in \mathcal{O}(e^{-\cdot^q})$ et $\mathcal{G}$ est $\|\mathcal{G}\|_{lip}$-**Lipschitz**, alors

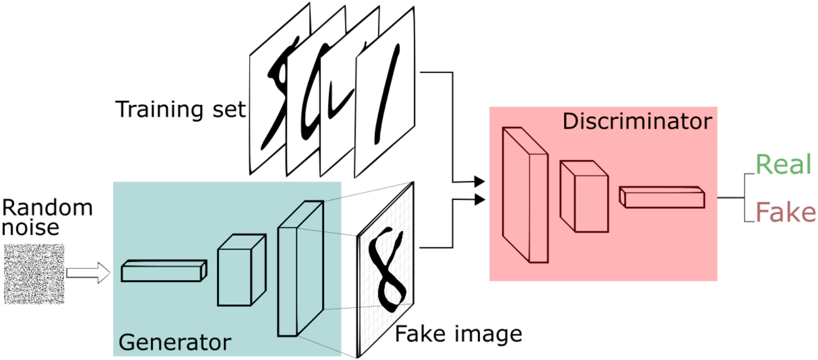
$$\mathcal{G}(X) \in \mathcal{O}\left(e^{-(\cdot/\|\mathcal{G}\|_{lip})^q}\right).$$

"Les vecteurs concentrés sont stable à travers des transformations Lipschitziennes."

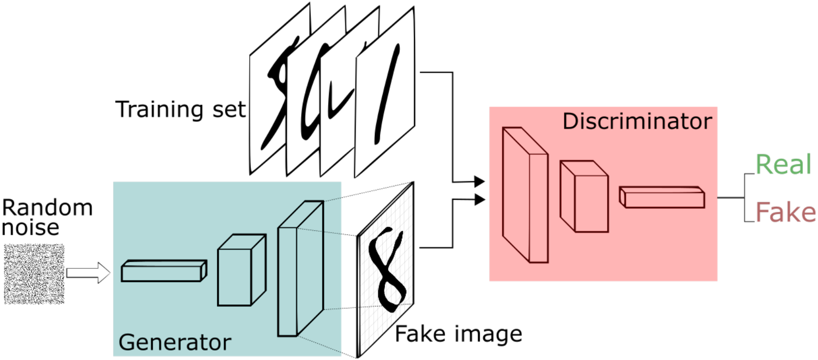
[1] **Reminder:** $\mathcal{F} : E \to F$ is $\|\mathcal{F}\|_{lip}$-Lipschitz if $\forall (x, y) \in E^2 : \|\mathcal{F}(x) - \mathcal{F}(y)\|_F \leq \|\mathcal{F}\|_{lip}\, \|x - y\|_E$.

Outline

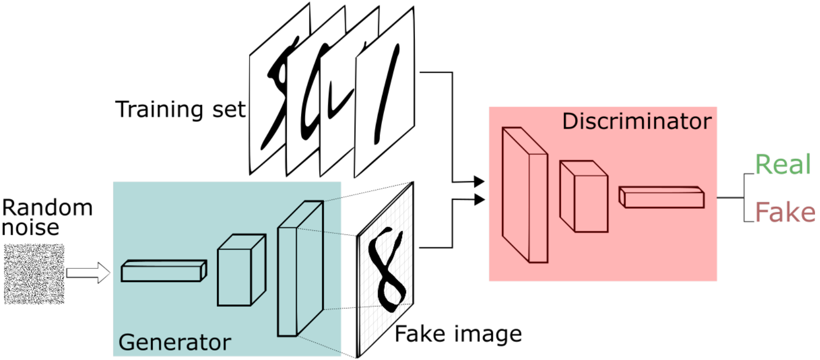
Données GAN : Un exemple de vecteurs concentrés



Données GAN : Un exemple de vecteurs concentrés



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



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

On génére les données comme

$$\boxed{\text{Image générée} = \mathcal{G}(\text{Gaussien})}$$

Données GAN : Un exemple de vecteurs concentrés

Figure: Images générées par le modèle BigGAN [Brock *et al*, ICLR'19].

Données GAN : Un exemple de vecteurs concentrés

Figure: Images générées par le modèle BigGAN [Brock *et al*, ICLR'19].

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

Données GAN : Un exemple de vecteurs concentrés

Figure: Images générées par le modèle BigGAN [Brock *et al*, ICLR'19].

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

où les $\mathcal{F}_i$'s sont des couches Fully Connected, Convolutionelles, Pooling et fonctions d'activation, des connextions résiduelles ou Batch Normalisation.

Données GAN : Un exemple de vecteurs concentrés

Figure: Images générées par le modèle BigGAN [Brock *et al*, ICLR'19].

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

où les $\mathcal{F}_i$'s sont des couches Fully Connected, Convolutionelles, Pooling et fonctions d'activation, des connextions résiduelles ou Batch Normalisation.

$\Rightarrow$ Les $\mathcal{F}_i$'s sont des opérations *Lipschitziennes*.

- **Couches Fully Connected Layers et Convolutionelles** sont des opérations affines :

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

et $\|\mathcal{F}_i\|_{lip} = \sup_{u \neq 0} \frac{\|W_i u\|_p}{\|u\|_p}$, pour toute p-norme.

- **Couches Fully Connected Layers et Convolutionelles** sont des opérations affines :

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

et $\|\mathcal{F}_i\|_{lip} = \sup_{u \neq 0} \frac{\|W_i u\|_p}{\|u\|_p}$, pour toute p-norme.

- **Pooling et Fonctions d'Activation :** Sont généralement au plus 1-Lipschitz par rapport à toute p-norme (e.g., ReLU and Max-pooling).

Données GAN : Un exemple de vecteurs concentrés

- **Couches Fully Connected Layers et Convolutionelles** sont des opérations affines :

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

et $\|\mathcal{F}_i\|_{lip} = \sup_{u \neq 0} \frac{\|W_i u\|_p}{\|u\|_p}$, pour toute p-norme.

- **Pooling et Fonctions d'Activation :** Sont généralement au plus 1-Lipschitz par rapport à toute p-norme (e.g., ReLU and Max-pooling).

- **Connexions résiduelles :** $\mathcal{F}_i(x) = x + \mathcal{F}_i^{(1)} \circ \cdots \circ \mathcal{F}_i^{(\ell)}(x)$
où les $\mathcal{F}_i^{(j)}$ sont des opérations Lipschitziennes, alors $\mathcal{F}_i$ est Lipschitz avec une constante de Lipschitz bornée par $1 + \prod_{j=1}^{\ell} \|\mathcal{F}_i^{(j)}\|_{lip}$.

- ...

Mixture of Concentrated Vectors

Considérons des données distribuées dans k classes $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ telles que

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

Considérons des données distribuées dans k classes $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ telles que

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

On note

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

Considérons des données distribuées dans k classes $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ telles que

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

On note

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

Hypothèses (Taux de croissance)

Quand $p \to \infty$,

Mixture of Concentrated Vectors

Considérons des données distribuées dans k classes $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ telles que

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

On note

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

Hypothèses (Taux de croissance)

Quand $p \to \infty$,

1. $p/n \to c \in (0, \infty)$.

Mixture of Concentrated Vectors

Considérons des données distribuées dans k classes $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ telles que

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

On note

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

Hypothèses (Taux de croissance)

Quand $p \to \infty$,

1. $p/n \to c \in (0, \infty)$.
2. *Le nombre de classes k est borné.*

Considérons des données distribuées dans k classes $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ telles que

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

On note

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

Hypothèses (Taux de croissance)

Quand $p \to \infty$,

1. $p/n \to c \in (0, \infty)$.
2. *Le nombre de classes k est borné.*
3. *Pour tout $\ell \in [k]$, $\|\mu_\ell\| = \mathcal{O}(\sqrt{p})$.*

Considérons des données distribuées dans k classes $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ telles que

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

On note

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

Hypothèses (Taux de croissance)

Quand $p \to \infty$,

1. $p/n \to c \in (0, \infty)$.
2. *Le nombre de classes k est borné.*
3. *Pour tout $\ell \in [k]$, $\|\mu_\ell\| = \mathcal{O}(\sqrt{p})$.*

Notation

$Q(z) = (X^\intercal X/p + zI_n)^{-1}$.

Outline

Comportement de la matrice de Gram de vecteurs concentrés

Soient

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

Soient

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

L la densité spectrale empirique de G et U la matrice qui contient les vecteurs propres dominants de G. Alors

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

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

Soient

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

L la densité spectrale empirique de G et U la matrice qui contient les vecteurs propres dominants de G. Alors

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

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

$\boxed{\Rightarrow \text{Analyser le comportement de la resolvent } Q(z) = (G + zI_n)^{-1}.}$

Théorème

Sous les hypothèses précédentes, on a $Q(z) \in \mathcal{O}(e^{-(\sqrt{p}\,\cdot)^q})$ *in* $(\mathbb{R}^{n\times n}, \|\cdot\|)$*. De plus,*

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

Théorème

Sous les hypothèses précédentes, on a $Q(z) \in \mathcal{O}(e^{-(\sqrt{p}\,\cdot)^q})$ *in* $(\mathbb{R}^{n\times n}, \|\cdot\|)$*. De plus,*

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

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

Théorème

Sous les hypothèses précédentes, on a $Q(z) \in \mathcal{O}(e^{-(\sqrt{p}\,\cdot)^q})$ *in* $(\mathbb{R}^{n\times n}, \|\cdot\|)$. *De plus,*

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

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

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

Théorème

Sous les hypothèses précédentes, on a $Q(z) \in \mathcal{O}(e^{-(\sqrt{p}\,\cdot)^q})$ *in* $(\mathbb{R}^{n\times n}, \|\cdot\|)$*. De plus,*

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

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

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

avec $\delta(z) = [\delta_1(z), \ldots, \delta_k(z)]$ *l'unique point fixe du système d'équations suivant*

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

Théorème

Sous les hypothèses précédentes, on a $Q(z) \in \mathcal{O}(e^{-(\sqrt{p}\,\cdot)^q})$ *in* $(\mathbb{R}^{n\times n}, \|\cdot\|)$. *De plus,*

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

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

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

avec $\delta(z) = [\delta_1(z), \ldots, \delta_k(z)]$ *l'unique point fixe du système d'équations suivant*

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

Théorème

Sous les hypothèses précédentes, on a $Q(z) \in \mathcal{O}(e^{-(\sqrt{p}\,\cdot)^q})$ *in* $(\mathbb{R}^{n\times n}, \|\cdot\|)$. *De plus,*

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

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

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

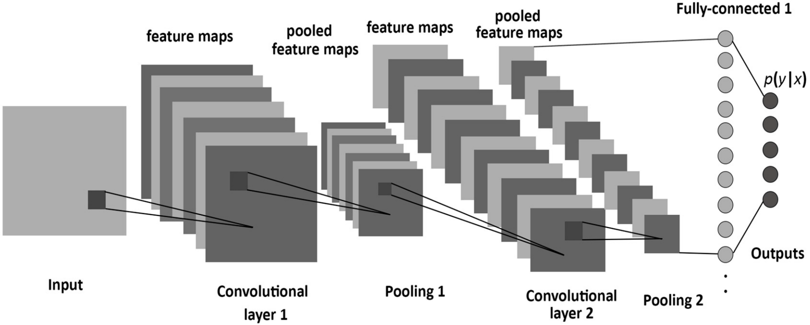
avec $\delta(z) = [\delta_1(z), \ldots, \delta_k(z)]$ *l'unique point fixe du système d'équations suivant*

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

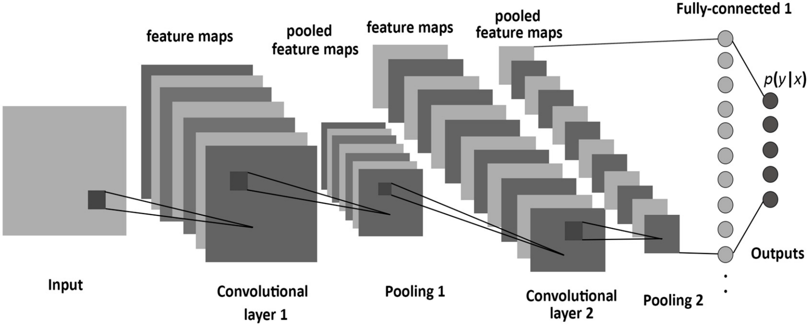
Observation clé : Seulement les moments d'ordre un et deux sont important !

Outline

Application aux representations CNN des images GAN

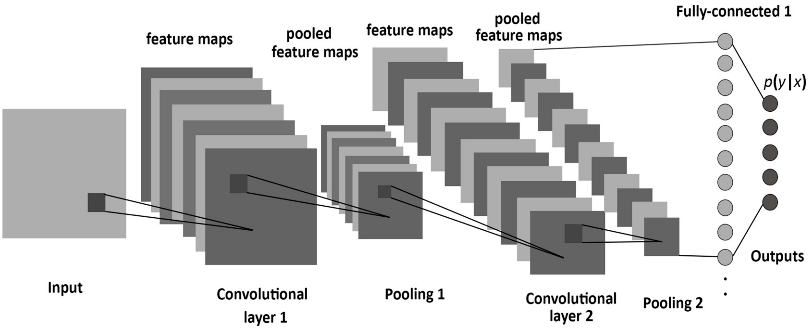


Application aux representations CNN des images GAN



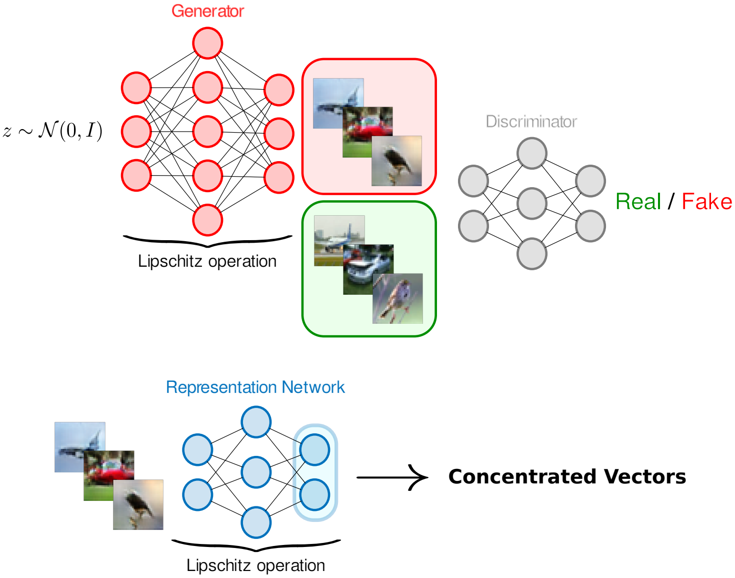
- Les représenations CNN correspondent à **l'avant dernière** couche du réseau.

Application aux representations CNN des images GAN



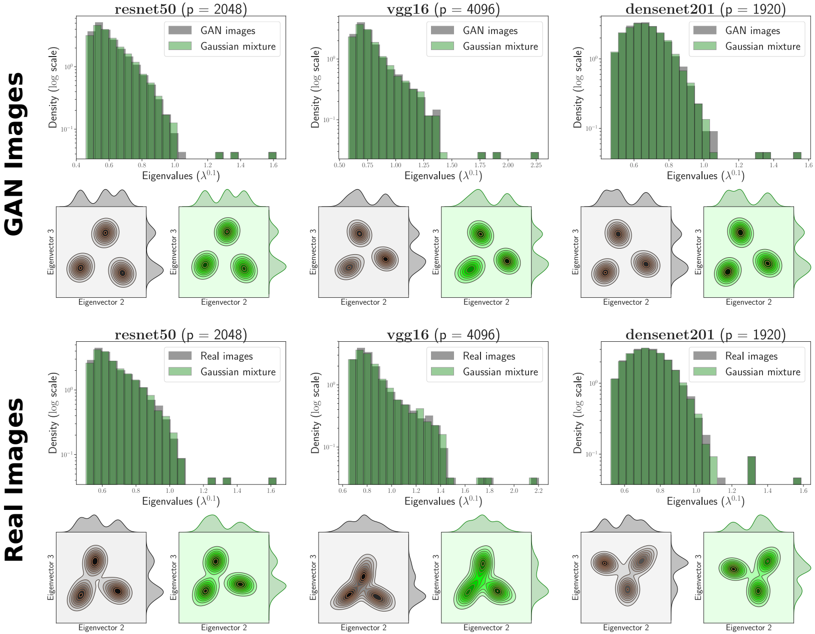
- Les représenations CNN correspondent à **l'avant dernière** couche du réseau.
- Architectures couramment utilisées en pratique : **Resnet, VGG, Densenet**.

Application aux representations CNN des images GAN




GAN Images
Real Images

Application aux representations CNN des images GAN



Merci pour votre attention !