

# WGAN-GP ET DISCRIMINATOR OPTIMAL TRANSPORT (DOT)

THÉORIE, MISE EN PRATIQUE SUR MNIST, ÉVALUATION, REGARD CRITIQUE  
N'HÉSITEZ PAS À ALLER SUR MON CANVA DIRECTEMENT (POUR LES GIF) :  
[HTTPS://WWW.CANVA.COM/DESIGN/DAGYKBUQXSS/3CEMB202MLF08VB9PKF](https://www.canva.com/design/DAGYKBUQXSS/3CEMB202MLF08VB9PKF)

UEA/EDIT?ID=EWIIJP7LKEJONRYDWV9FQ

# SOMMAIRE

## Théorie générale

### A Concepts clés

- Norme Lipschitz
- Distance de Wasserstein

### B Generative Adversarial Network

- GAN Classique
- Wasserstein GAN(WGAN)

### C WGAN-Gradient Penalty (WGANGP)

## Le transport optimal pour les GAN

### A Le transport optimal

- Pourquoi le transport optimal ?
- La théorie

### B Application pour les WGAN

- Calcul de la norme de lipschitz en pratique
- Target Space DOT
- Latent Space DOT

### C Présentation des résultats

- visualisation des données
- visualisation des effets sur les distribution

## Conclusion

### A regard critique sur les 2 méthodes

### B Comment l'améliorer?

# CONCEPTS CLÉS

## Norme Lipschitz

$$\|f\|_{\text{Lip}} := \sup \left\{ \frac{\|f(x) - f(y)\|_2}{\|x - y\|_2} \mid x \neq y \right\}$$

- Taux de variation **maximal** d'une fonction  $f$ .
- Quantifie à quel point une petite modification de l'entrée  $x$  peut provoquer une variation dans la sortie  $f(x)$
- Avoir une norme de Lipschitz trop élevée ou faible implique souvent **instabilité**

- **Coût minimal** pour transformer une distribution  $p$  en une autre  $q$ .
- Prend en compte la **géométrie des données** :
- Reste bien définie même si les distributions  $p$  et  $q$  n'ont pas de support commun

## distance de Wasserstein

$$W(p, q) = \inf_{\gamma \in \Pi(p, q)} \int \|x - y\| d\gamma(x, y)$$

$$W(p, q) = \inf_{\gamma \in \Pi(p, q)} \mathbb{E}_{(X, Y) \sim \gamma} [\|X - Y\|]$$

# LES GANS CLASSIQUES : DES LIMITES THÉORIQUES ET PRATIQUES

01

JENSEN-SHANON DIVERGENCE

02

INSTABILITÉ DE L'ENTRAÎNEMENT

03

VANISHING GRADIENT

04

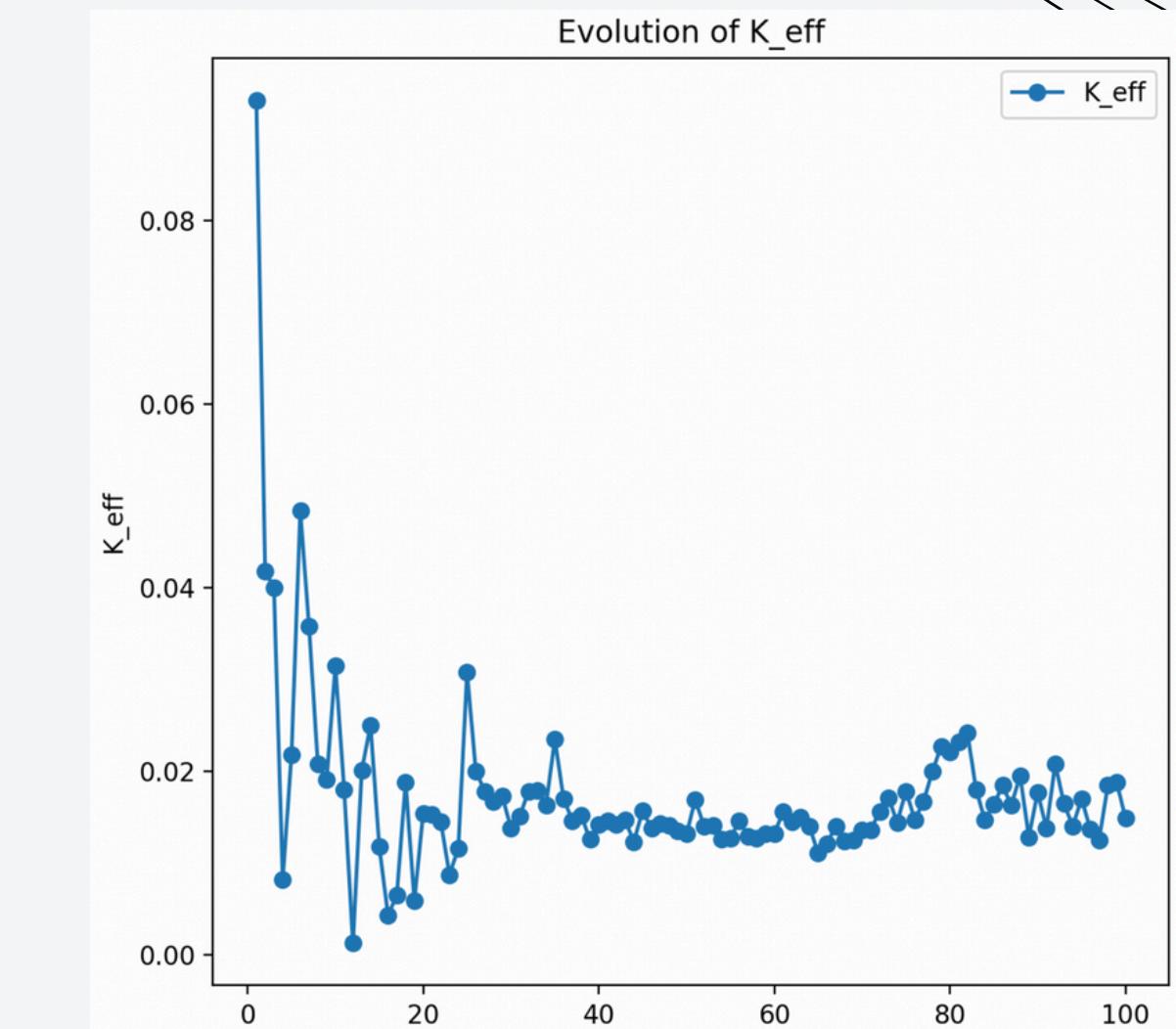
MODE COLLAPSE

05

DIFFICULTÉ DE CALCULER LA  
NORME LIPSCHITZ

06

DIFFICULTÉ POUR LE TRANSPORT  
OPTIMAL



$$V(G, D) = \mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log (1 - D(G(z)))] .$$

$$V(G, D^*) = -\log(4) + 2 \cdot JS(p_{\text{data}} \| p_G),$$

$$D^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_G(x)}.$$

# WGAN

théorème de Kantorovich–Rubinsteïn

$$[W(P, Q) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x \sim P}[f(x)] - \mathbb{E}_{x \sim Q}[f(x)]]$$

*Loss du Générateur*

$$L_G = -\mathbb{E}_{x \sim Q}[D(x)]$$

*Loss du Discriminateur*

$$L_D = \mathbb{E}_{x \sim Q}[D(x)] - \mathbb{E}_{x \sim P}[D(x)]$$

Comment respecter la contrainte Lipschitz?

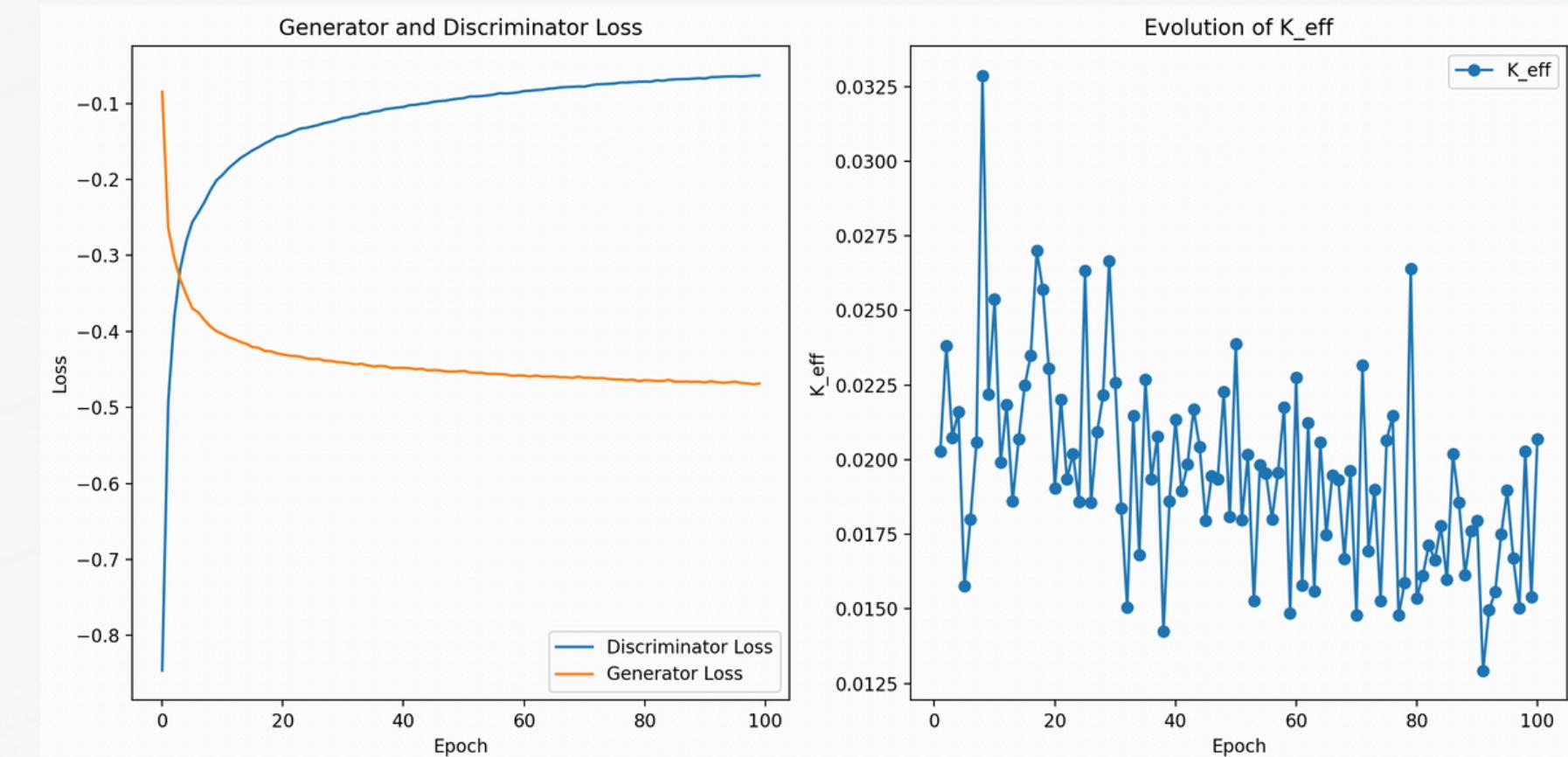
- Une méthode : normaliser avec la norme spectrale

$$SN(l_{W,b}) = l_{\left(\frac{W}{\sigma(W)}, b\right)}$$

- s'appuie sur cette propriété :  $\|f \circ g\|_{\text{Lip}} \leq \|f\|_{\text{Lip}} \cdot \|g\|_{\text{Lip}}$

- qui nous assure que :  $\|f_{nn}\|_{\text{Lip}} \leq \prod_{l=1}^L \left\| l_{\frac{W_l}{\sigma(W_l)}} \right\|_{\text{Lip}} = 1$

- Manière simple de calculer la distance de Wasserstein.
- $D(x; \theta)$  approxime la fonction  $f$ .



- Entrainement stable mais extrêmement lent
- pas de réel contrôle sur la norme Lipschitz

# LE GRADIENT PENALTY

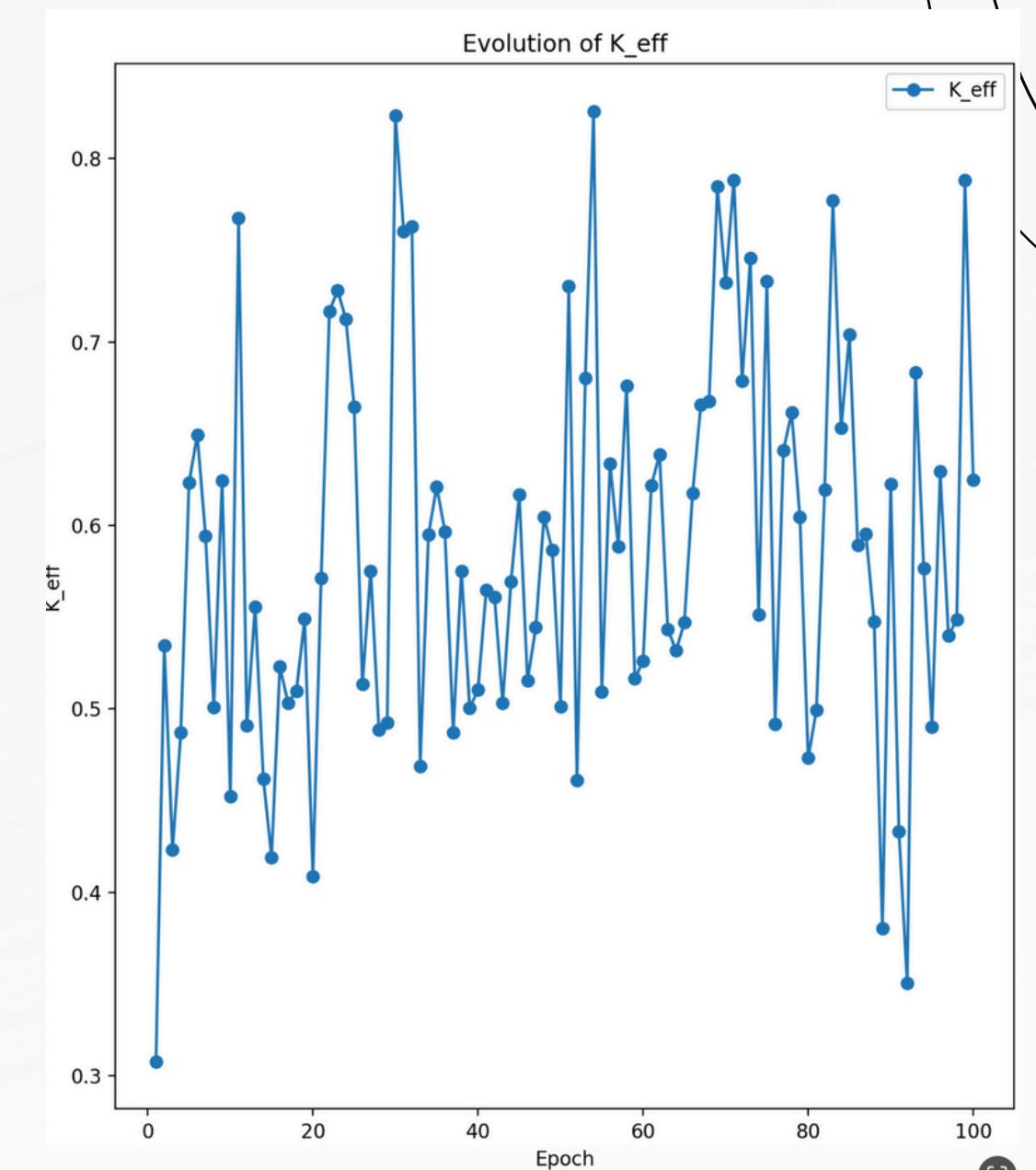
- une alternative à la norme spectrale
- on fait plus que contraindre, on demande à la norme de converger vers 1

$$GP = \mathbb{E}_{\hat{x} \sim \mathbb{P}_{\hat{x}}} [(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2]$$

**Nouvelle loss :**  $L_D = \mathbb{E}_{x \sim Q}[D(x)] - \mathbb{E}_{x \sim P}[D(x)] + \lambda \cdot \mathbb{E}_{\hat{x} \sim \mathbb{P}_{\hat{x}}} [(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2]$

- Méthode plus douce que le gradient clipping
- Utile pour DOT d'avoir une norme 1 pour le discriminatuer

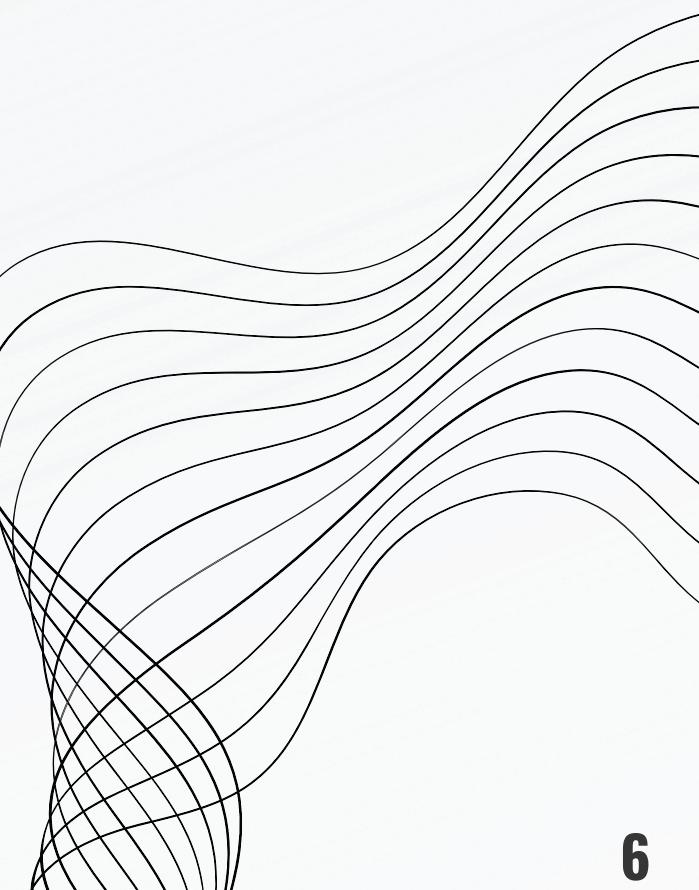
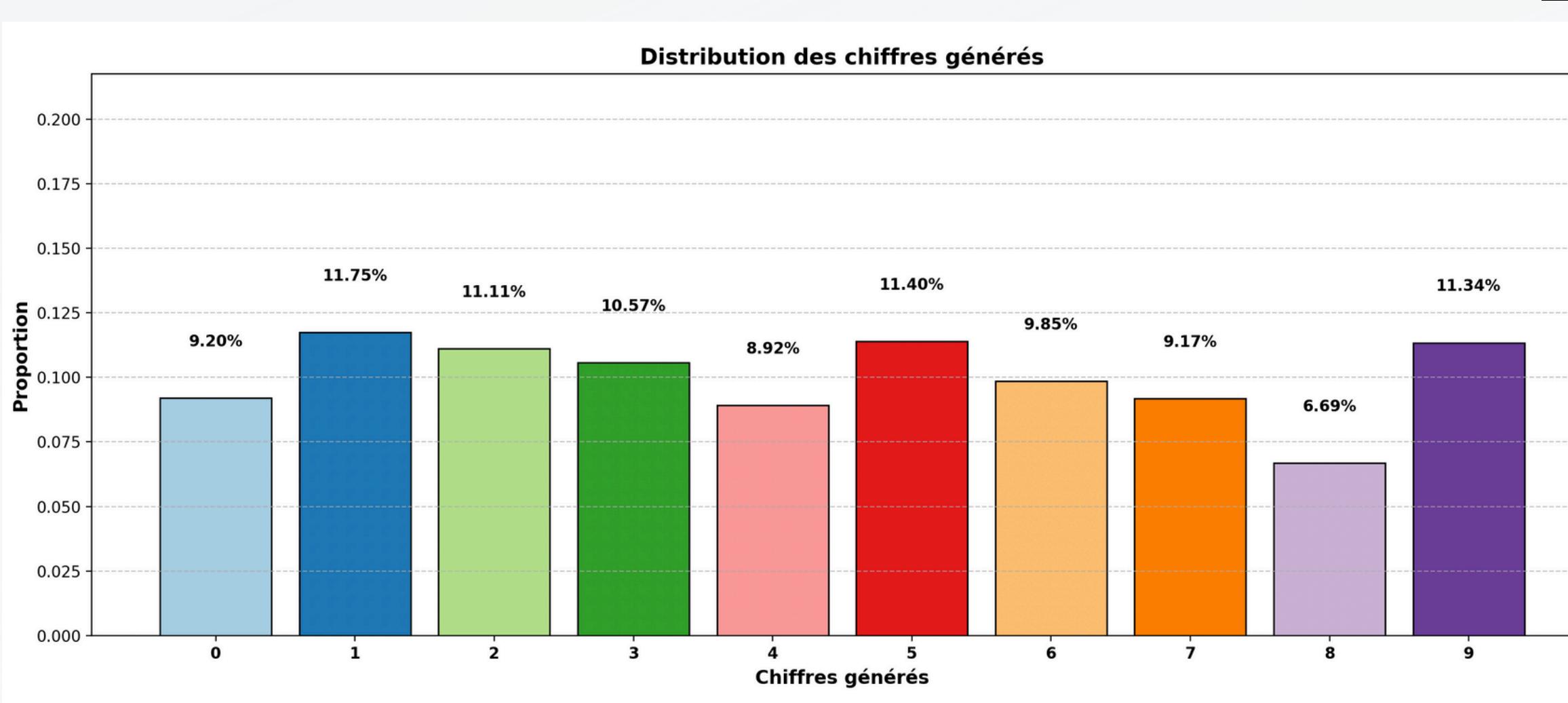
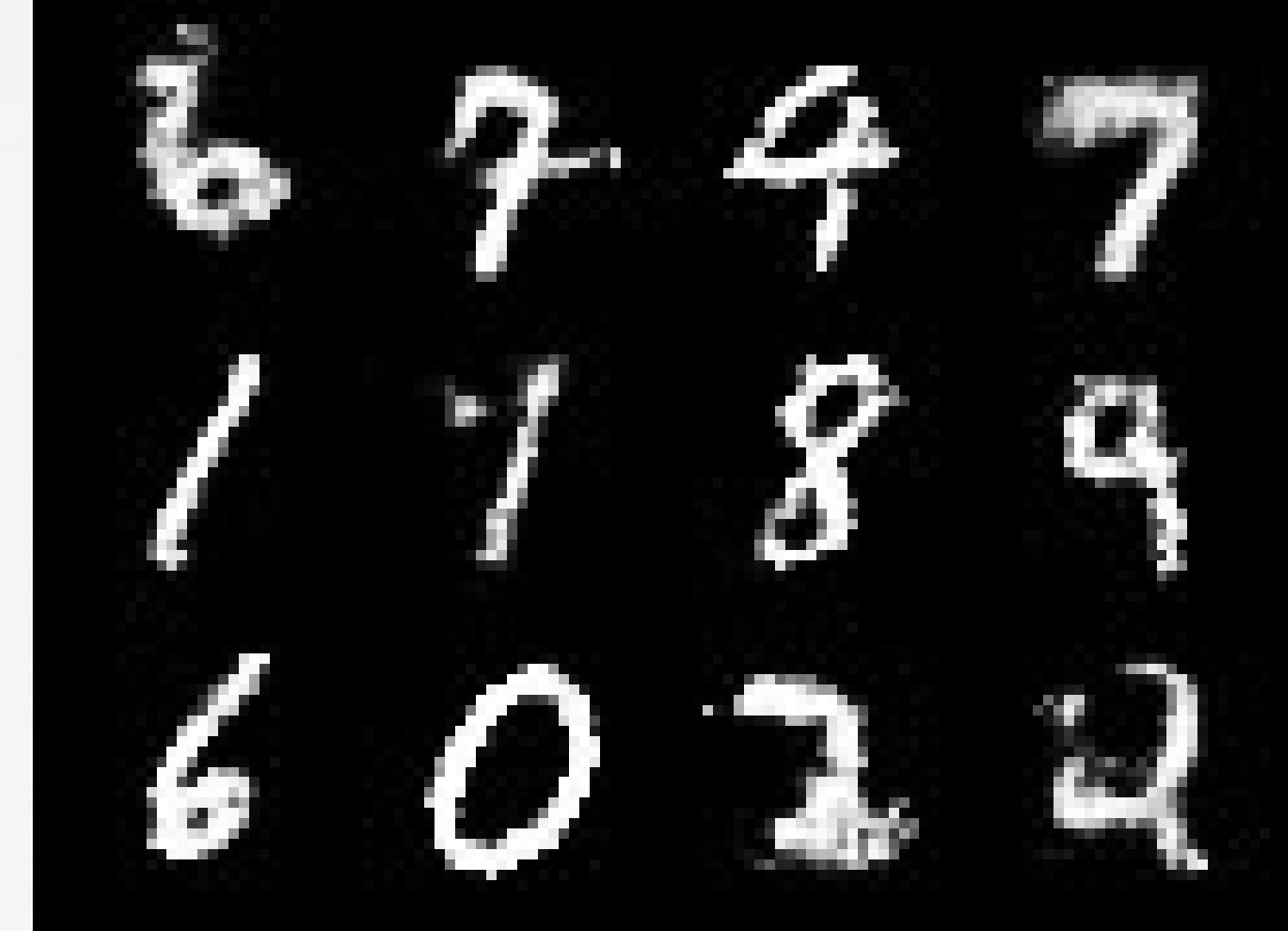
Les variations sont attendues :  
c'est une estimation empirique

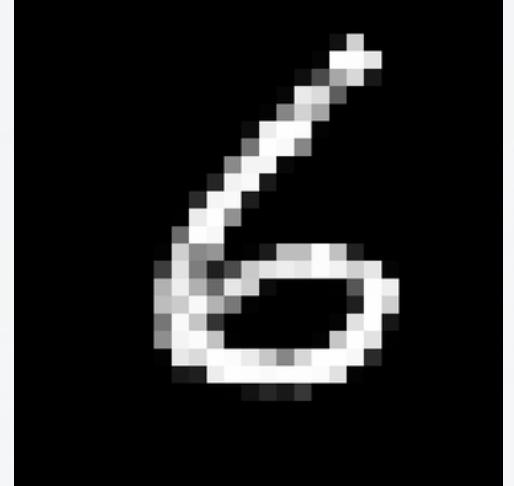


# WGAN-GP

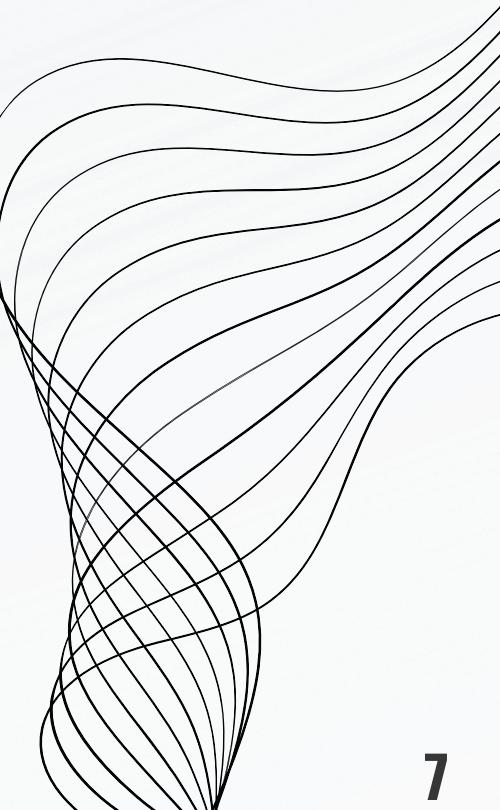
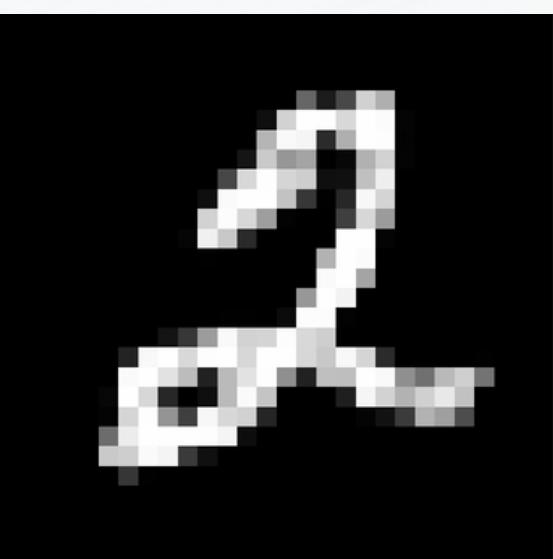
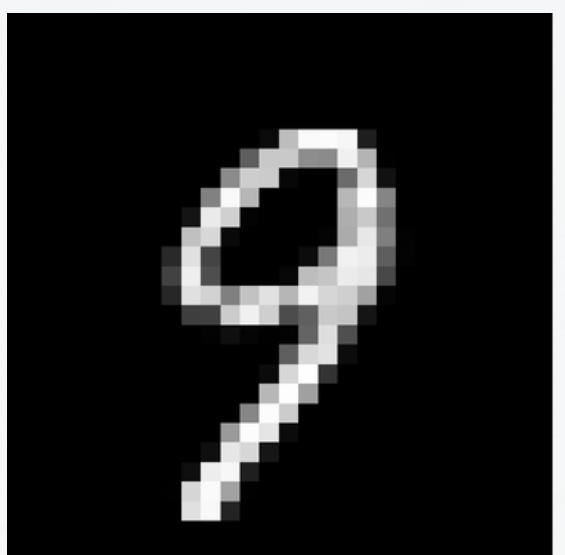
Au final

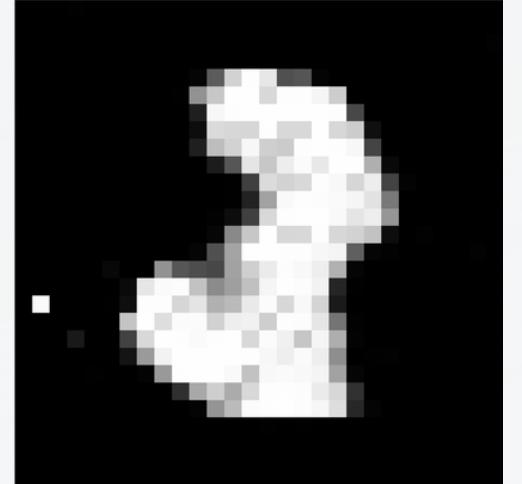
- Modèle concluant, sans encore avoir implémenté de méthodes post-train
- Entrainement très stable
- Pas de mode collapse en vu, au contraire une extrême diversité des chiffres



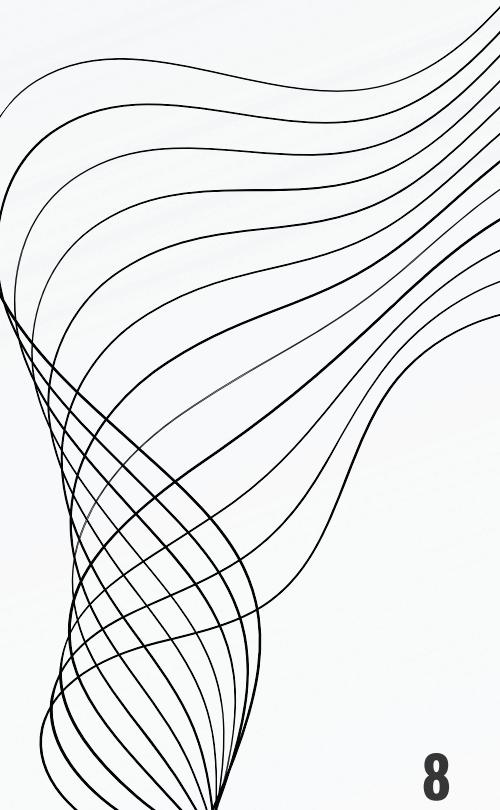
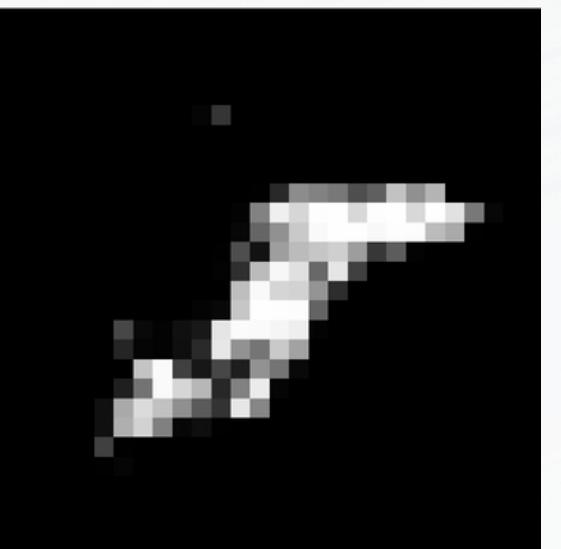
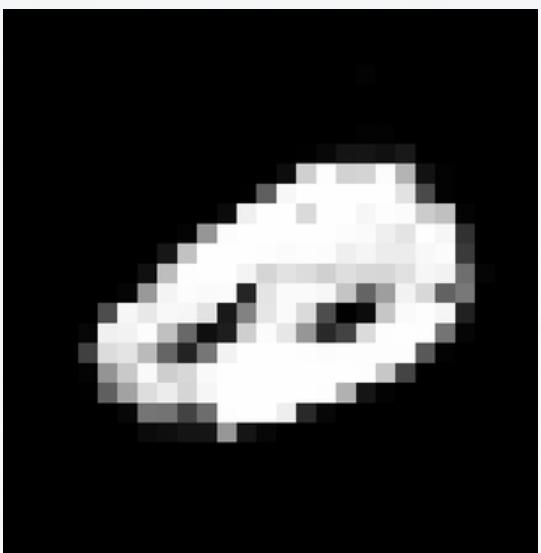


# LE TRANSPORT OPTIMAL

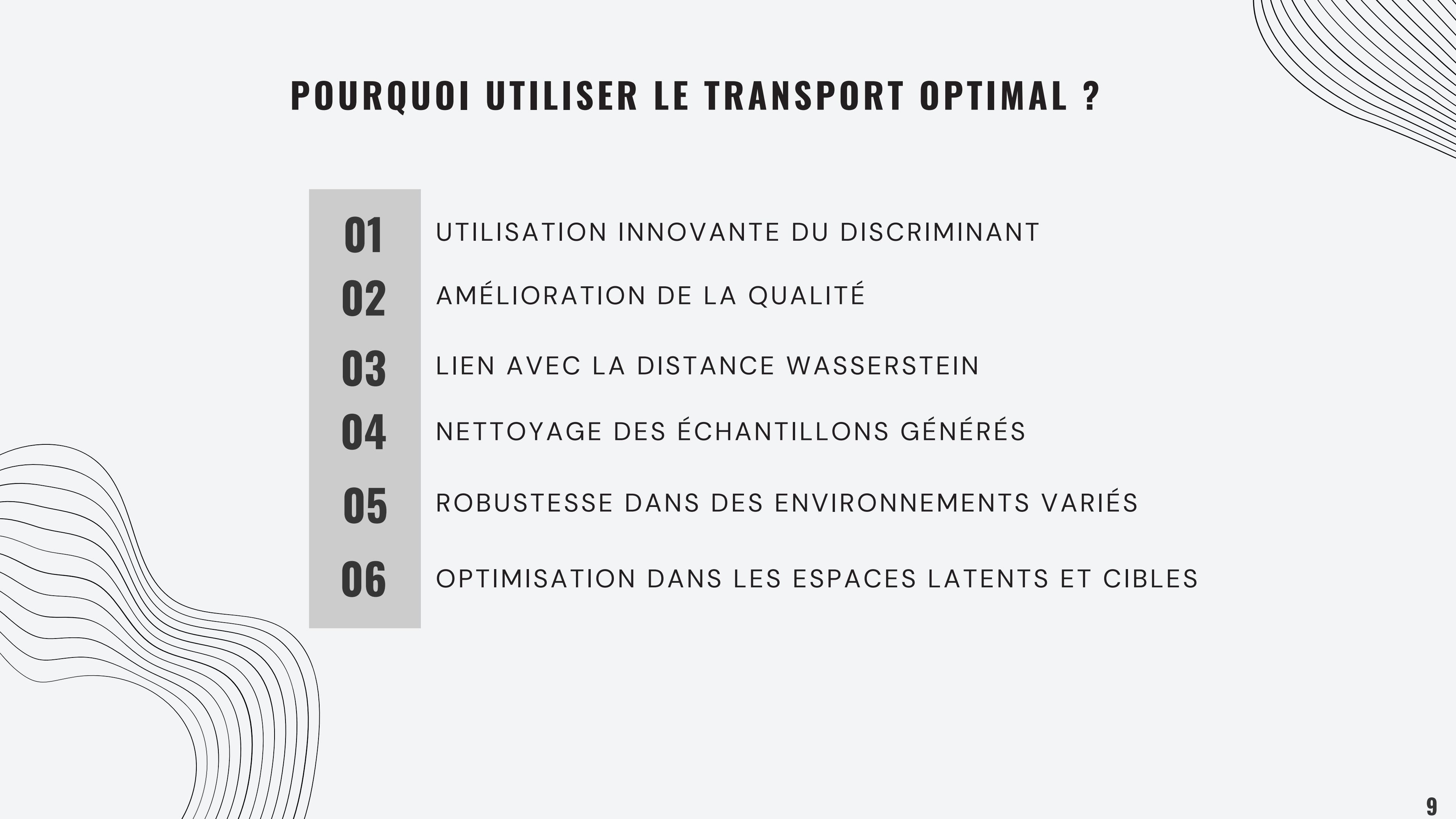




# LE TRANSPORT OPTIMAL



# POURQUOI UTILISER LE TRANSPORT OPTIMAL ?

- 
- 01** UTILISATION INNOVANTE DU DISCRIMINANT
  - 02** AMÉLIORATION DE LA QUALITÉ
  - 03** LIEN AVEC LA DISTANCE WASSERSTEIN
  - 04** NETTOYAGE DES ÉCHANTILLONS GÉNÉRÉS
  - 05** ROBUSTESSE DANS DES ENVIRONNEMENTS VARIÉS
  - 06** OPTIMISATION DANS LES ESPACES LATENTS ET CIBLES

# TRANSPORT OPTIMAL

*Deux théorèmes fondamentaux :*

Supposons que  $\pi^*$  et  $D^*$  soient des solutions optimales.  
Si  $\pi^*$  est un transport optimal déterministe décrit par un certain automorphisme  $T : X \rightarrow X$

- $\|D^*\|_{\text{Lip}} = 1$
- $T(y) = \arg \min_x \{\|x - y\|_2 - D^*(x)\}$
- $p(x) = \int dy \delta(x - T(y))q(y)$

Les deux théorèmes soulignent :

- le **pont** entre transport optimal et discriminant GAN
- Le lien entre maximiser la loss du discriminateur et **minimiser la distance de Wasserstein**
- L'intérêt de calculer **précisément** le coefficient Lipschitz

Chaque fonction objective de GAN avec une pénalité sur le gradient, fournit une borne inférieure de la divergence moyenne de  $\tilde{D} = D/K$  entre  $p$  et  $p_G$

$$V_D(G, D) \leq K \left( \mathbb{E}_{x \sim p} [\tilde{D}(x)] - \mathbb{E}_{y \sim p_G} [\tilde{D}(y)] \right)$$

*Discriminator Optimal Transport  
(version idéal)*

$$T_D(y) = \arg \min_x \left\{ \|x - y\|_2^2 - \frac{1}{K} D(x) \right\}$$

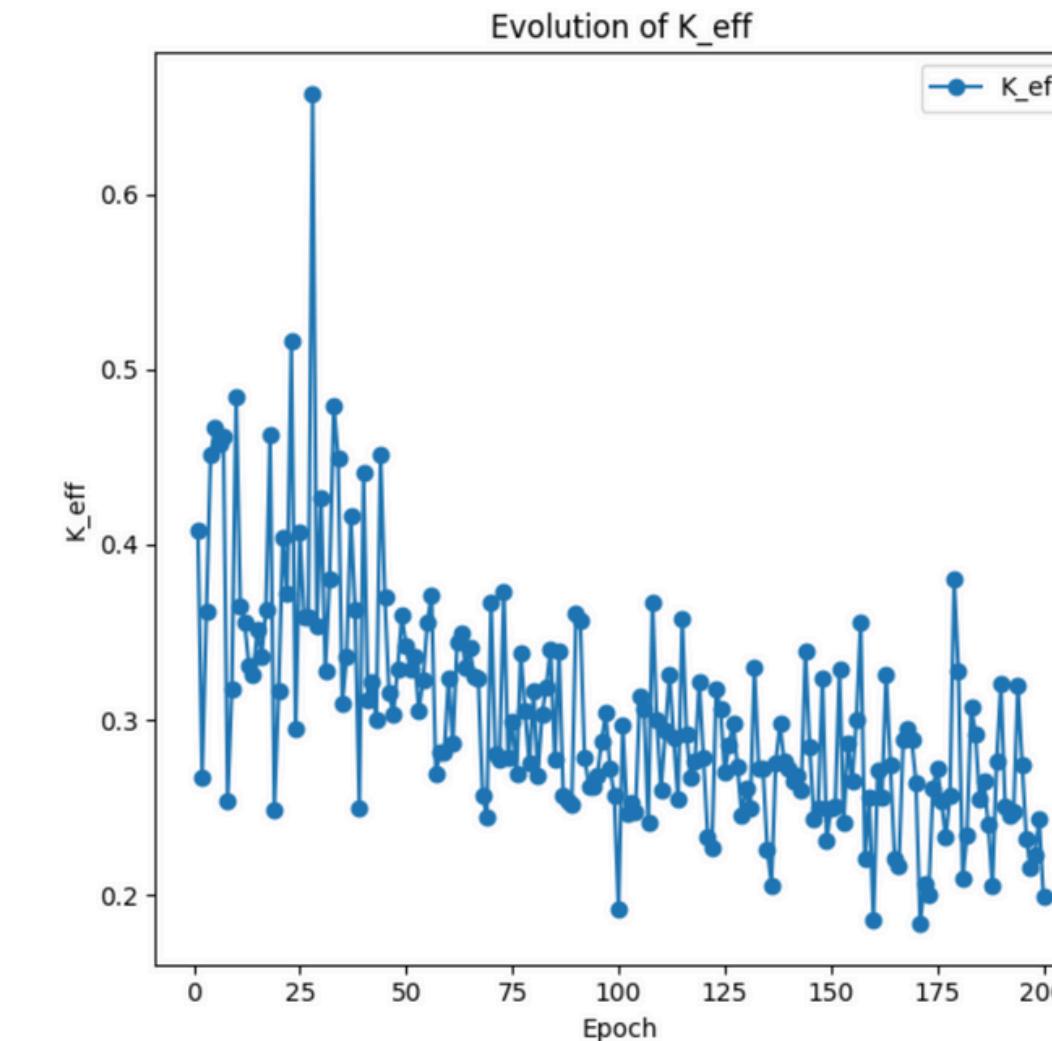
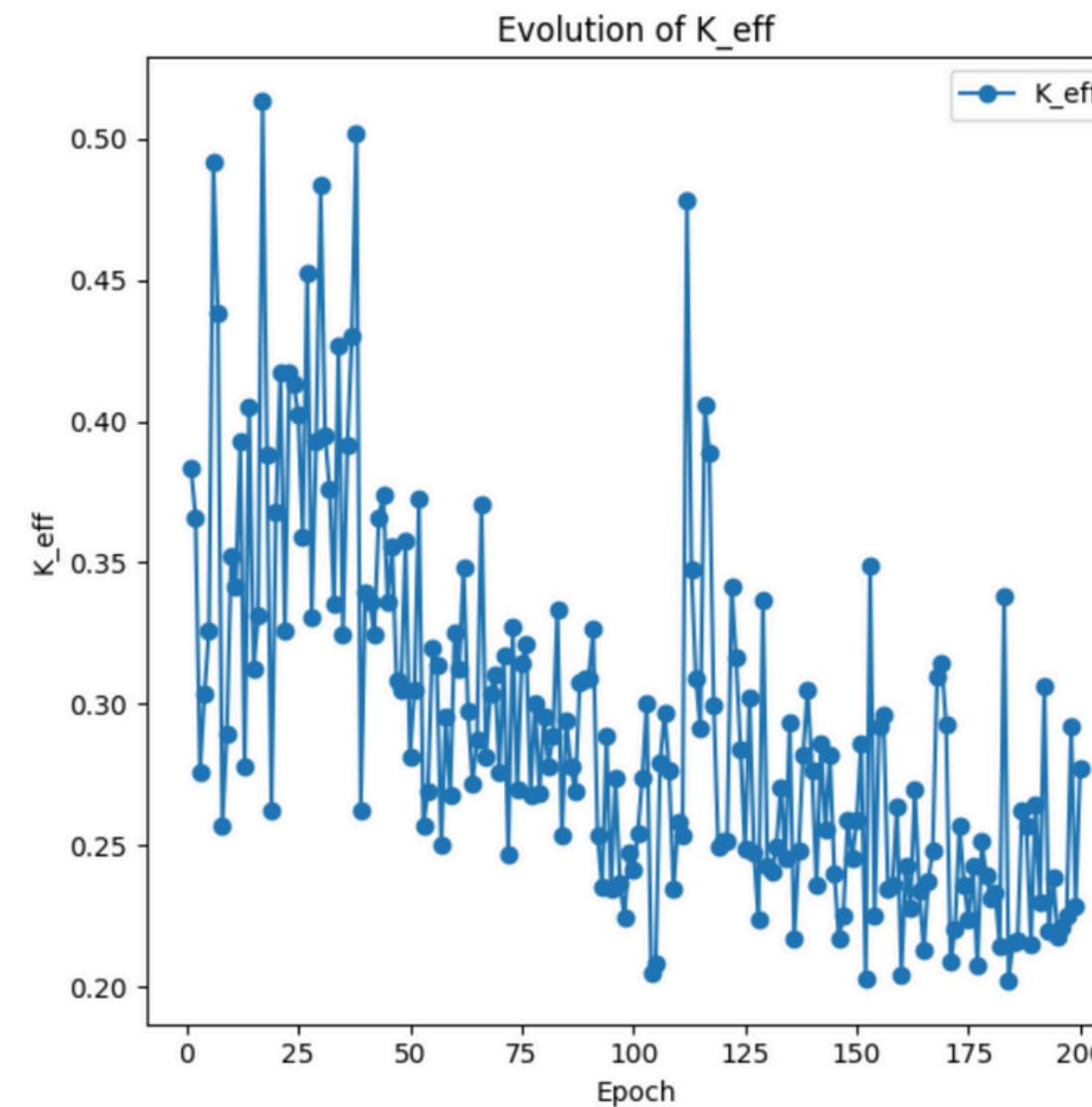
# LA PRATIQUE

## Calcul du coefficient lipschitz

$$K_{\text{eff}} = \max \left\{ \frac{|D(x) - D(y)|}{\|x - y\|_2} \mid x, y \sim p_G \right\}$$

$$k_{\text{eff}} = \max \left\{ \frac{|D \circ G(z) - D \circ G(z')|}{\|z - z'\|_2} \mid z, z' \sim p_Z \right\}$$

- Essentiel pour appliquer DOT
- permet de s'assurer que le discriminant respecte la contrainte de Lipschitz.
- stabilité pendant l'entraînement



# TARGET OPTIMAL TRANSPORT

## Transport optimal dans l'espace des images

$$T_D^{\text{eff}}(y) = \arg \min_x \left\{ \|x - y\|_2^2 - \frac{1}{K_{\text{eff}}} D(x) \right\}.$$

---

**Algorithm 1** Target space optimal transport by gradient descent

**Require:** trained  $D$ , approximated  $K_{\text{eff}}$  by (20), sample  $y$ , learning rate  $\epsilon$  and small vector  $\delta$

Initialize  $x \leftarrow y$

**for**  $n_{\text{trial}}$  in range( $N_{\text{updates}}$ ) **do**

$x \leftarrow x - \epsilon \nabla_x \left\{ \|x - y + \delta\|_2^2 - \frac{1}{K_{\text{eff}}} D(x) \right\}$  (  $\delta$  is for preventing overflow. )

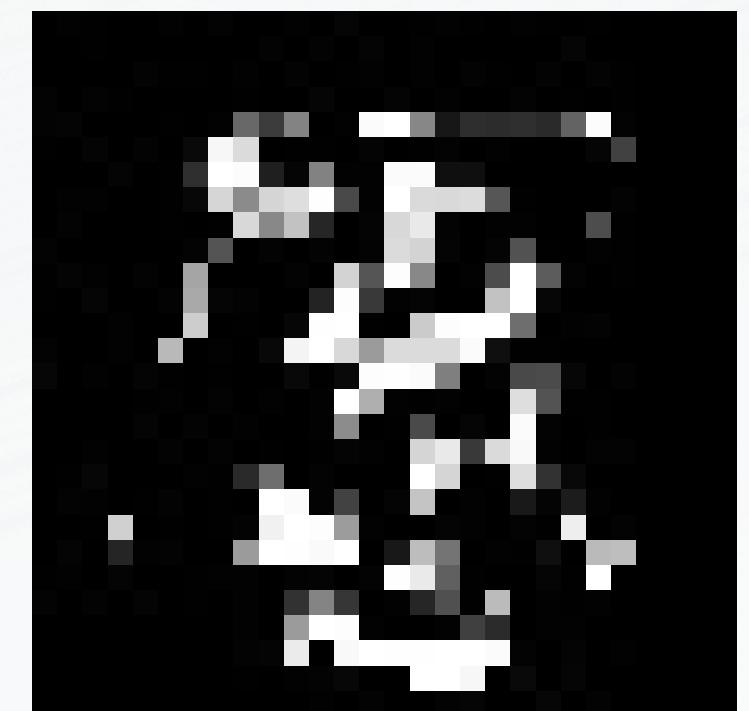
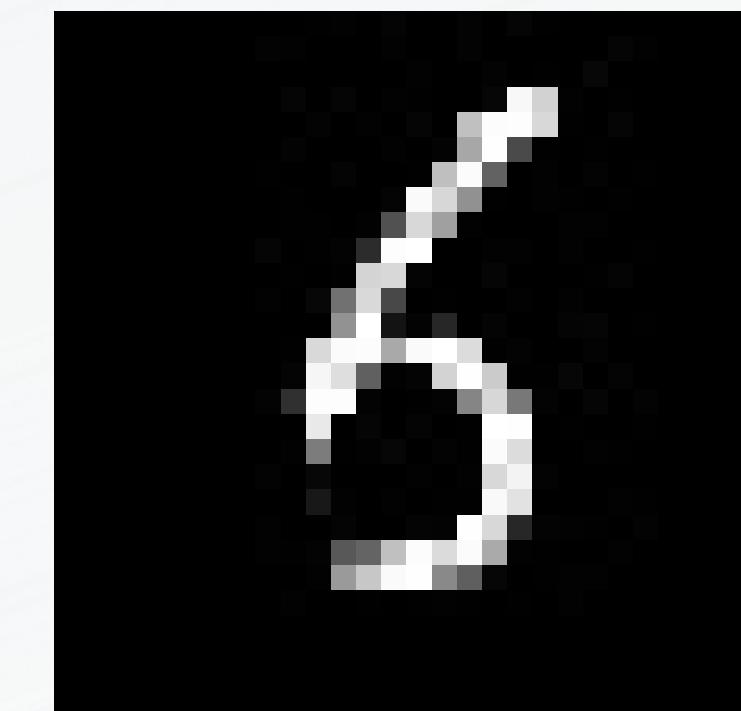
**end for**

**return**  $x$

---

Résultat :

- Amélioration minime
- Fonctionne pour des petites dimensions, utilité très restreinte avec mnist de dimension 784.
- Très instable



# LATENT SPACE DOT

## Transport optimal dans l'espace latent

$$T_{D \circ G}^{\text{eff}}(z_y) = \arg \min_z \left\{ \|z - z_y\|_2^2 - \frac{1}{k_{\text{eff}}} D \circ G(z) \right\}$$

Résultats :

- Changements non négligeables
- sensibilité très forte aux paramètres
- risque de mode collapse

**Algorithm 2** Latent space optimal transport by gradient descent

**Require:** trained  $G$  and  $D$ , approximated  $k_{\text{eff}}$ , sample  $z_y$ , learning rate  $\epsilon$ , and small vector  $\delta$

Initialize  $z \leftarrow z_y$

**for**  $n_{\text{trial}}$  in range( $N_{\text{updates}}$ ) **do**

$g = \nabla_z \left\{ \|z - z_y + \delta\|_2^2 - \frac{1}{k_{\text{eff}}} D \circ G(z) \right\}$  (  $\delta$  is for preventing overflow. )

**if** noise is generated by  $\mathcal{N}(0, I_{D \times D})$  **then**

$g \leftarrow g - (g \cdot z)z/\sqrt{D}$

**end if**

$z \leftarrow z - \epsilon g$

**if** noise is generated by  $U([-1, 1])$  **then**

        clip  $z \in [-1, 1]$

**end if**

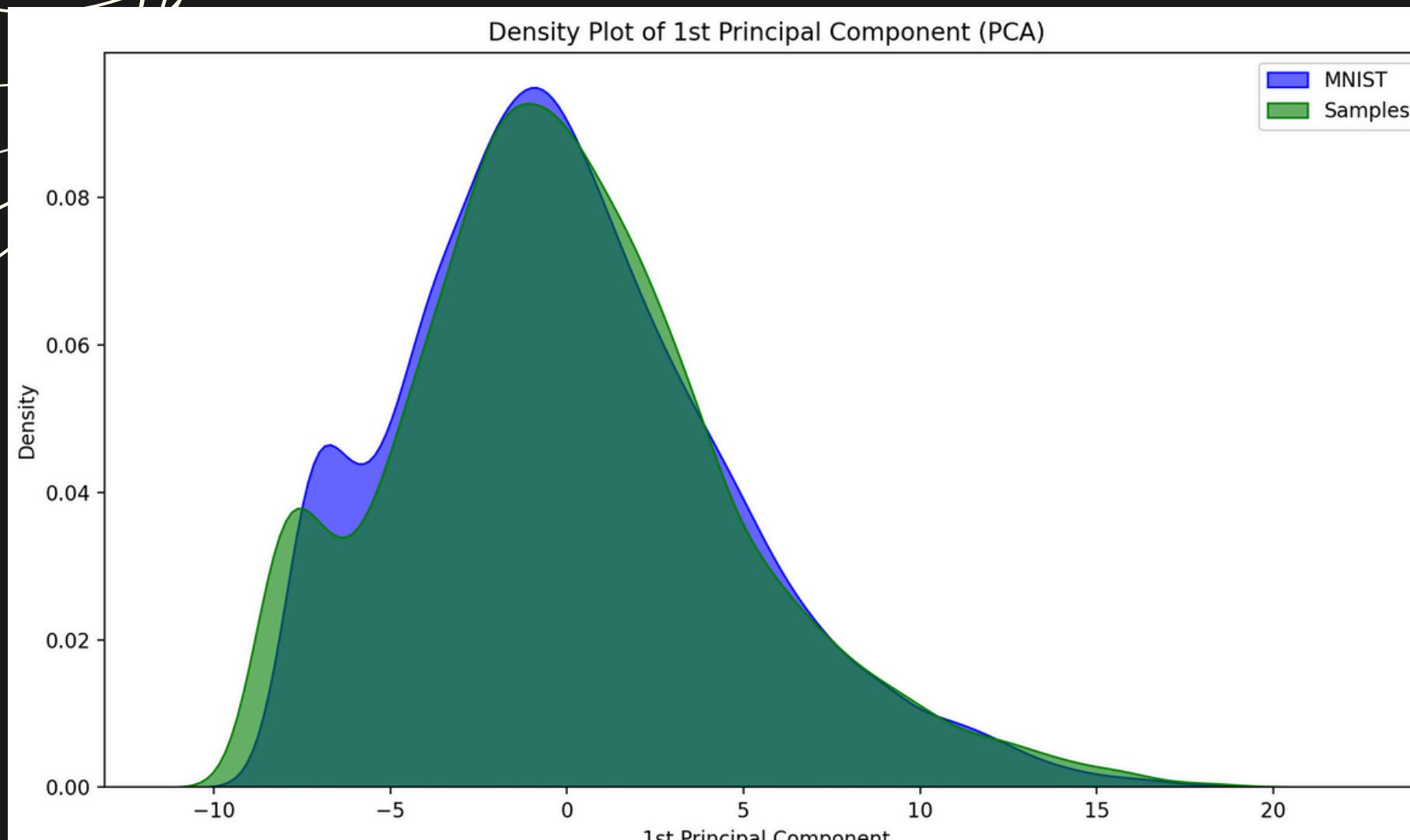
**end for**

**return**  $x = G(z)$

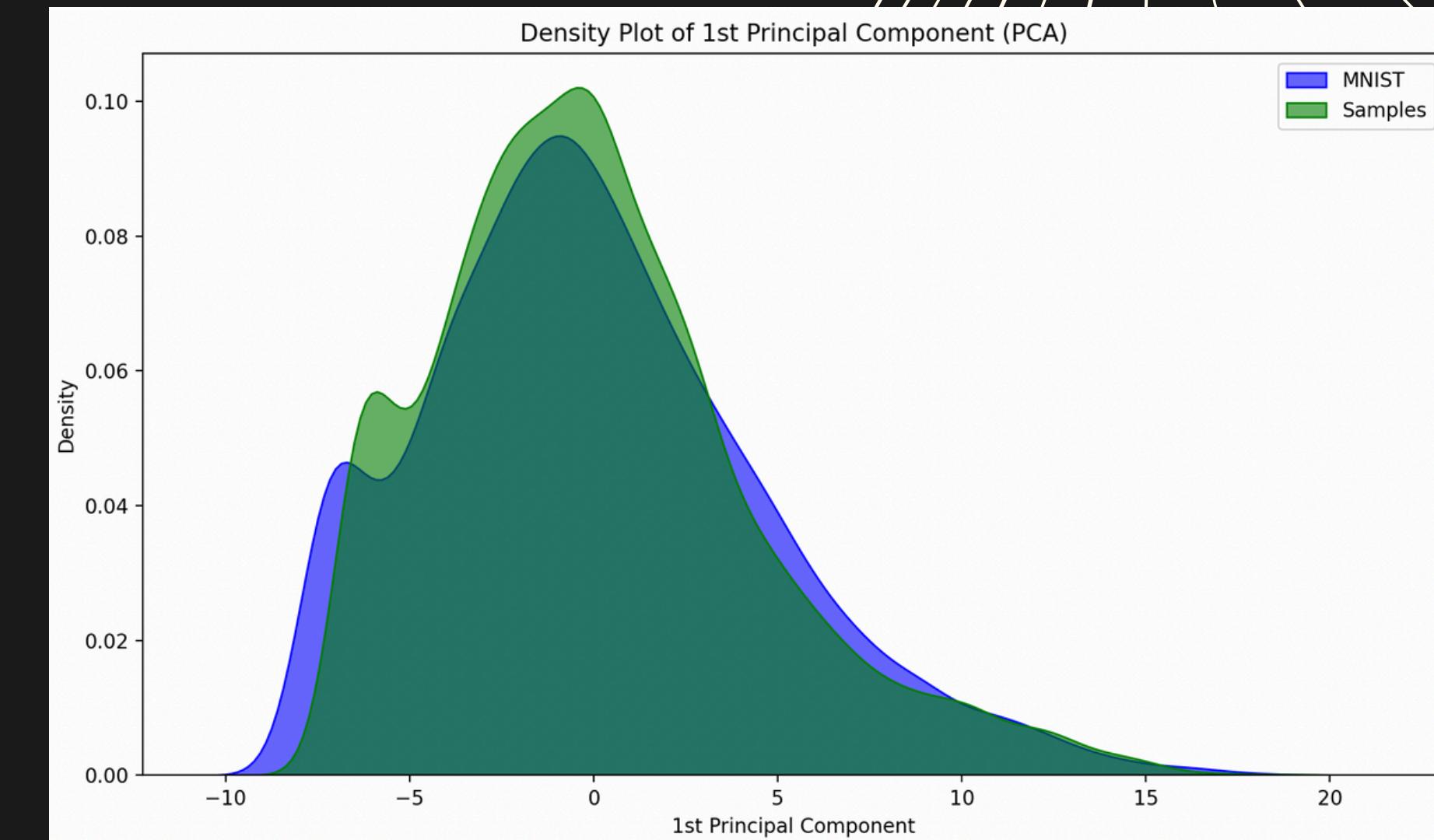


# Arbitrage Precision recall

WGANGP

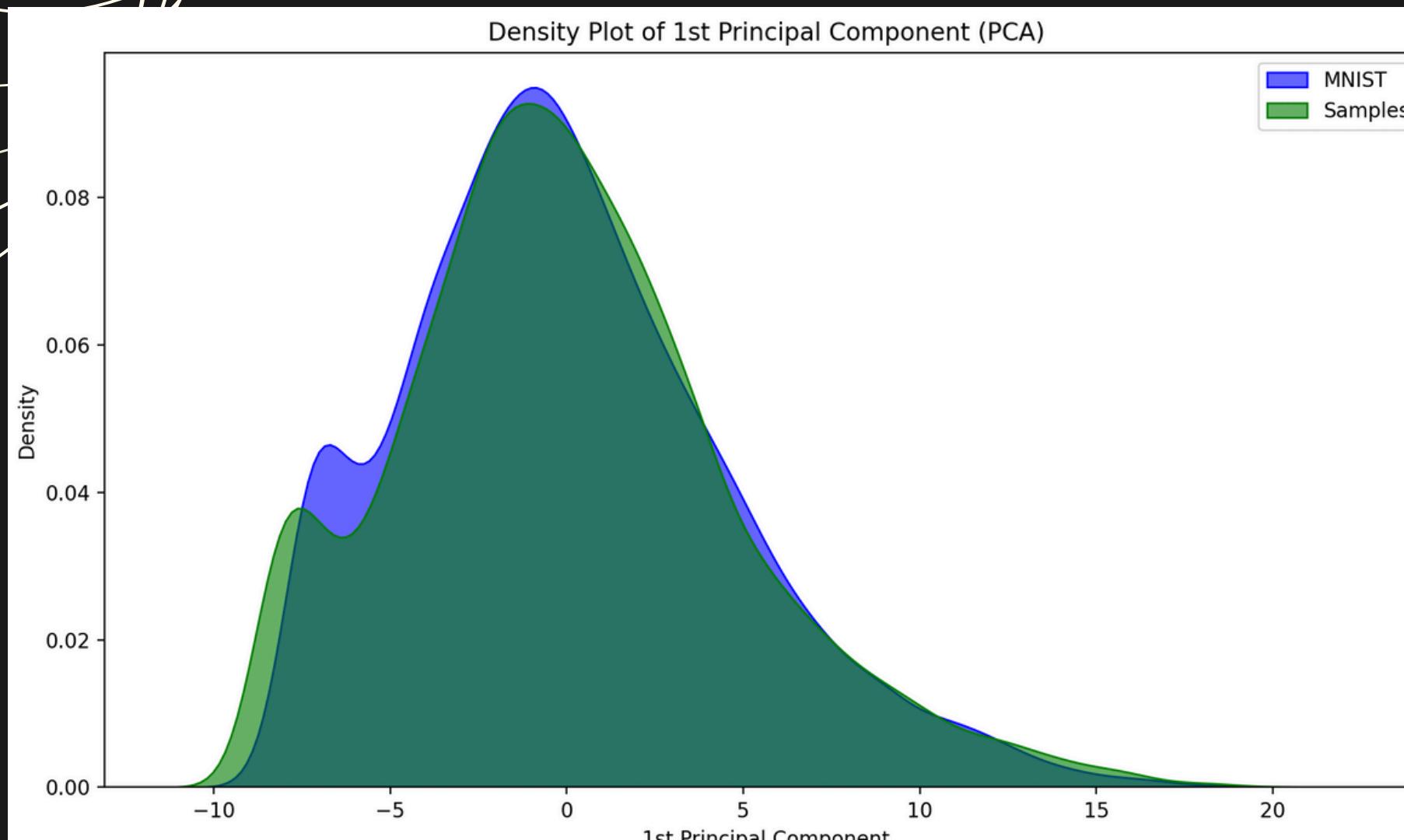


WGANGP + DOT

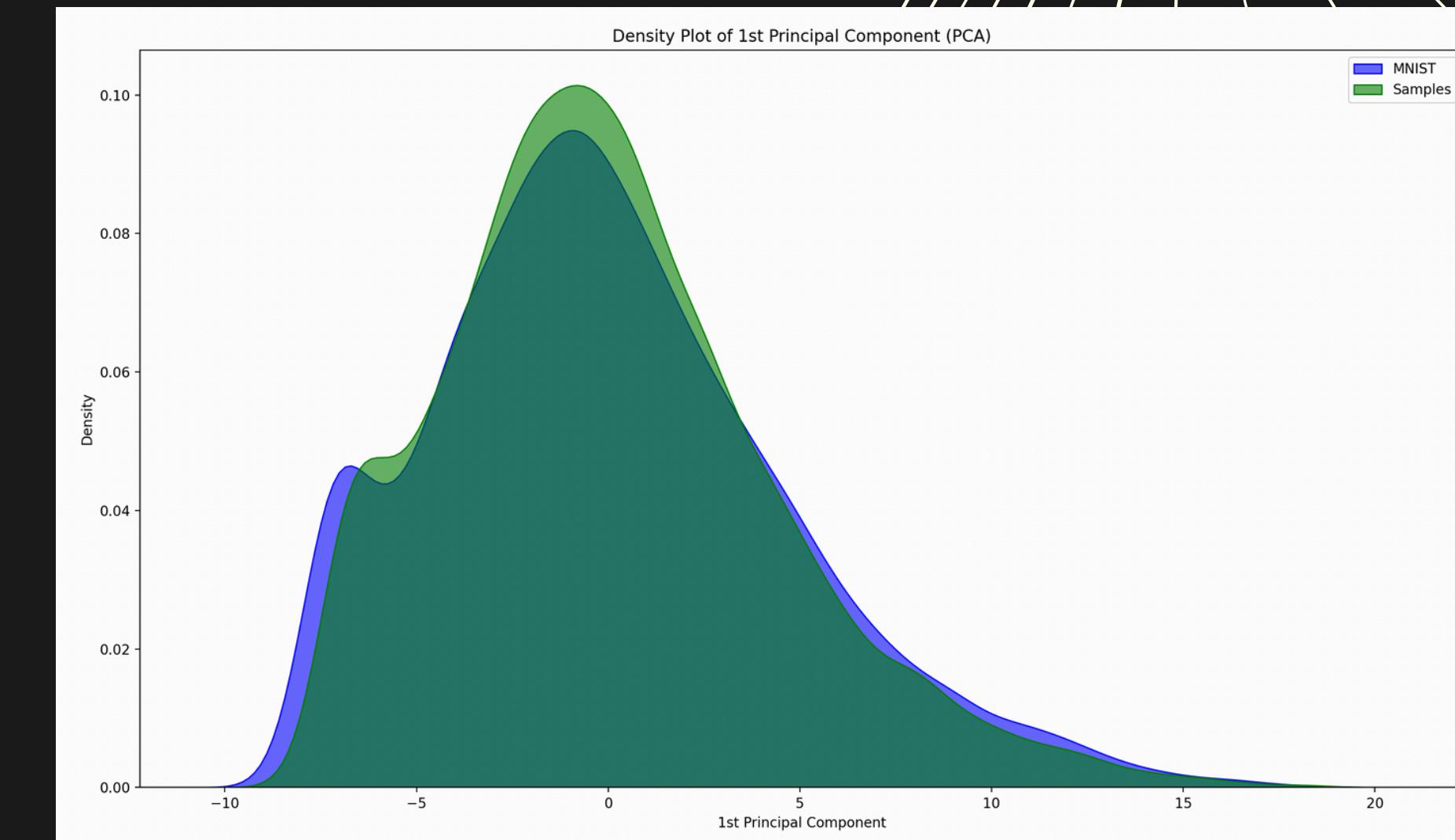


# Arbitrage Precision recall

WGANGP



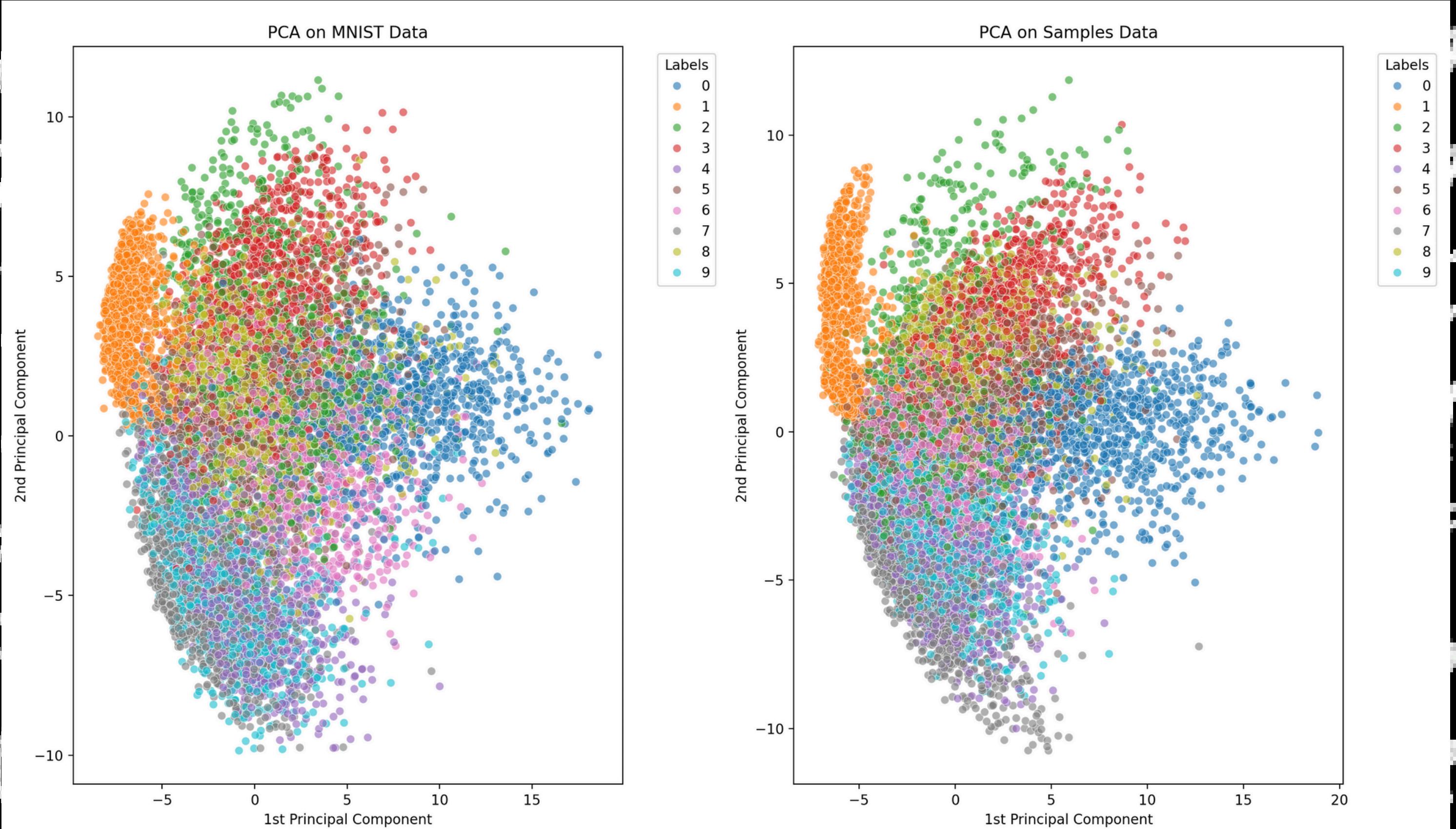
WGANGP + DOT



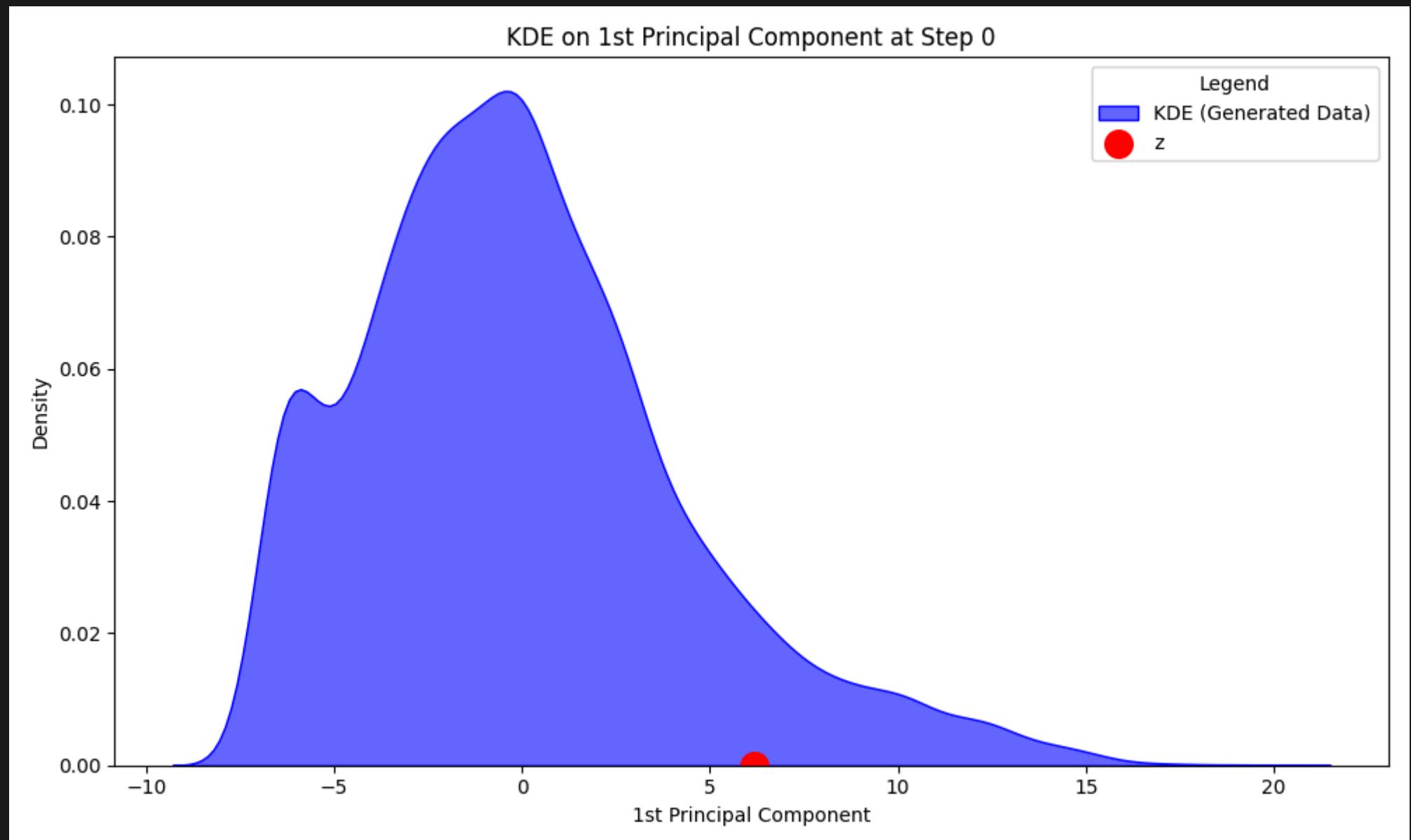
# WGAN



# WGAN + DOT

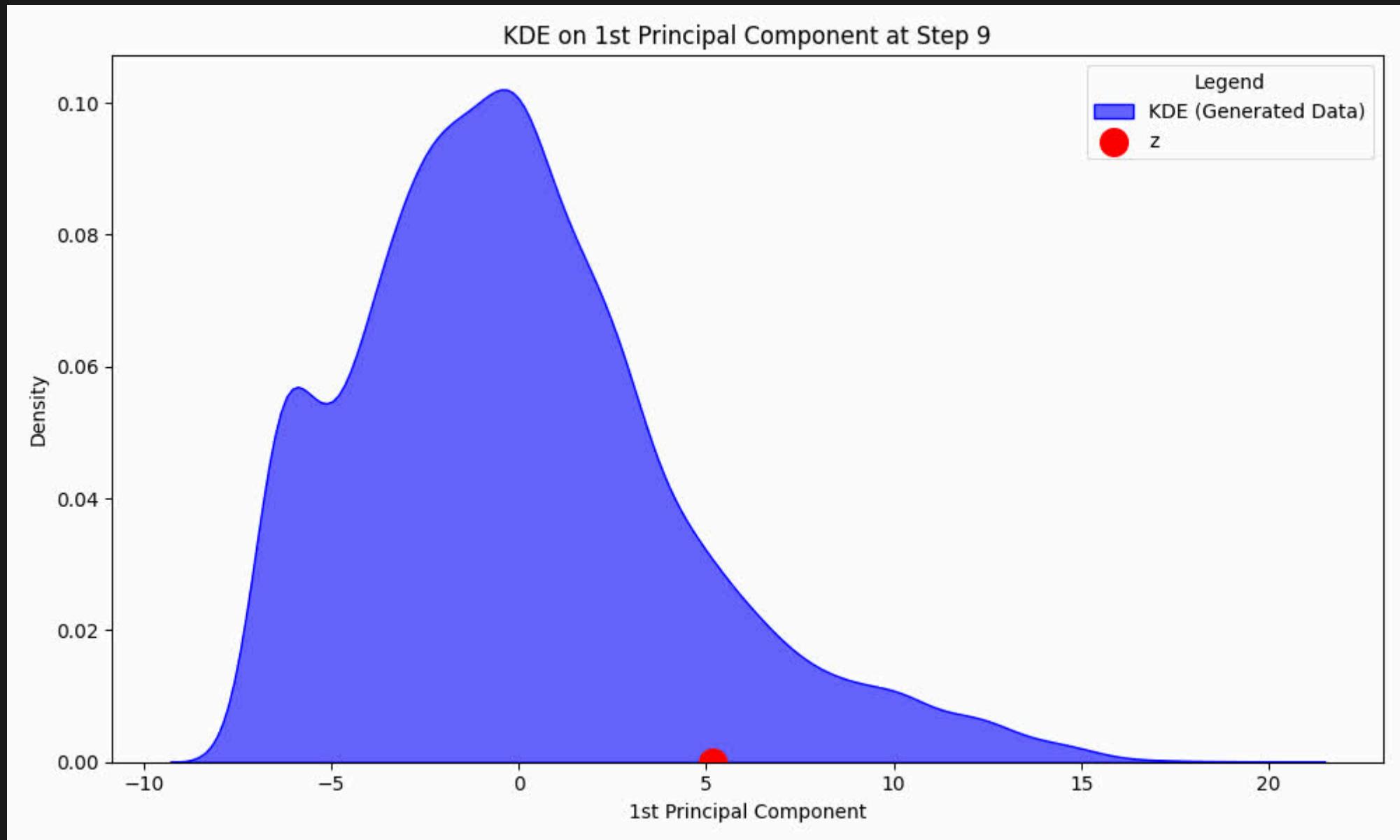


# Arbitrage Precision recall



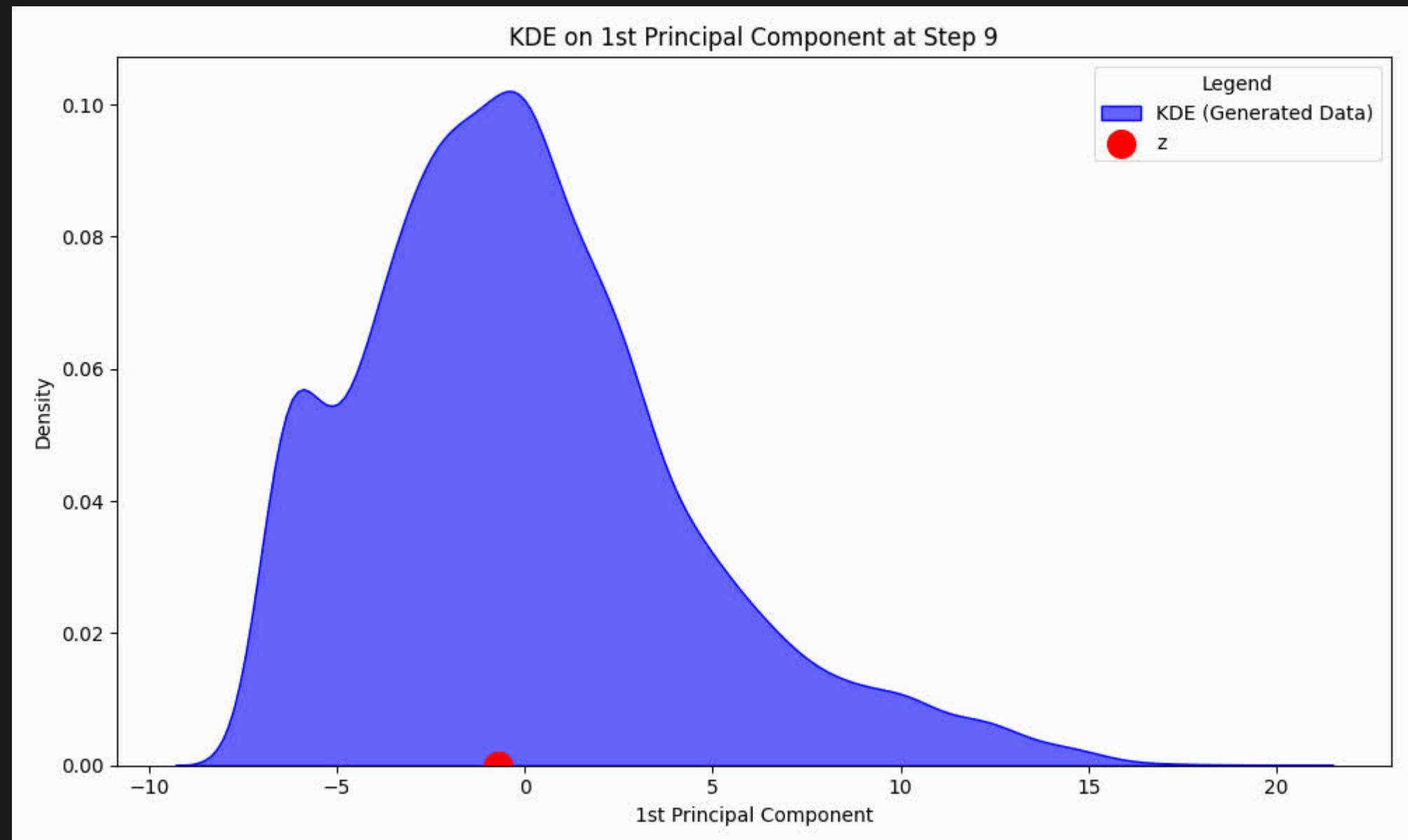
On optimise les données directement dans l'espace latent, gros risque de collapse et de réduire le recall, mais amélioration de la précision

# Arbitrage Precision recall

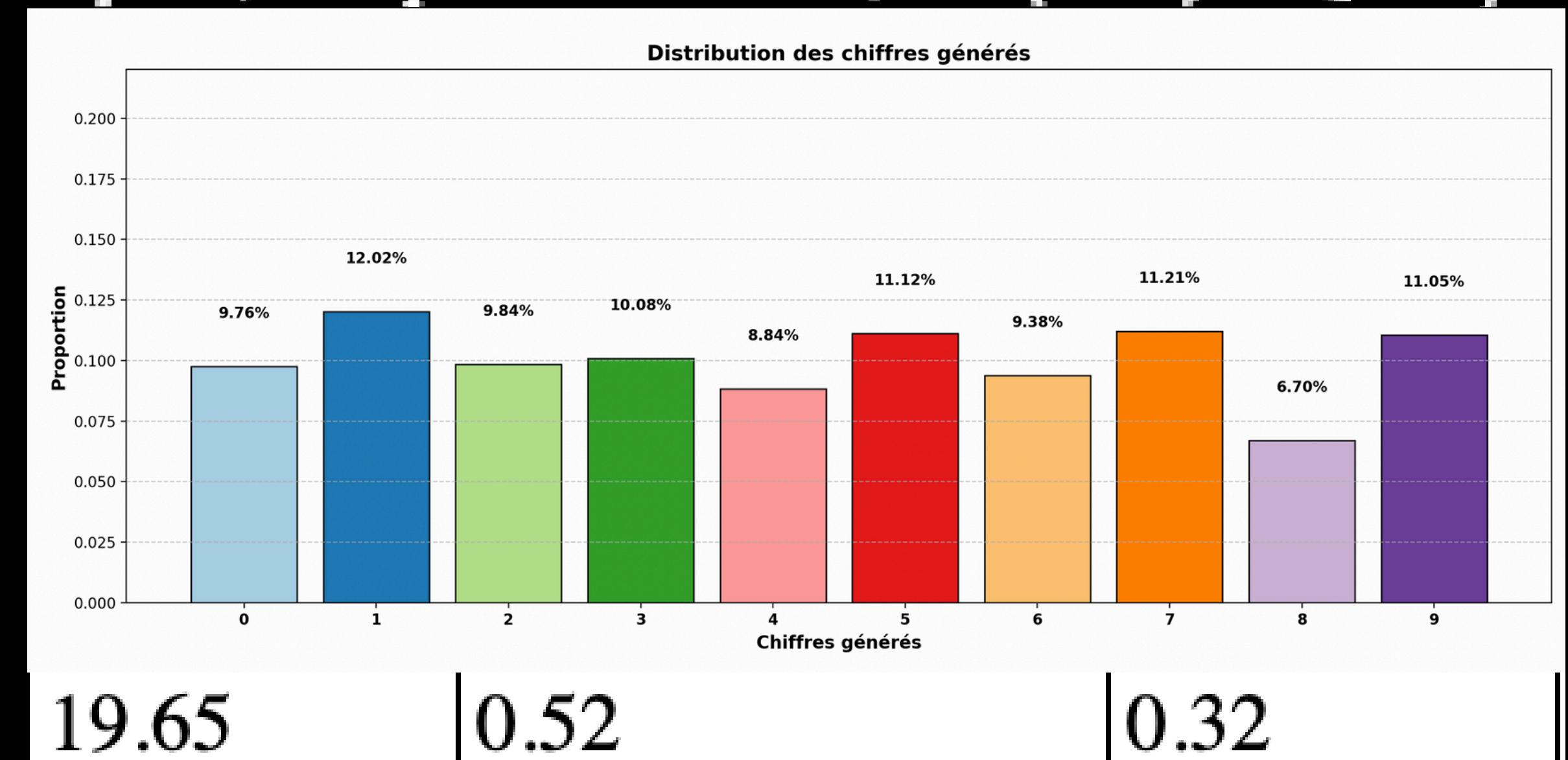


On optimise les données directement dans l'espace latent, gros risque de collapse et de réduire le recall, mais amélioration de la précision

# Arbitrage Precision recall



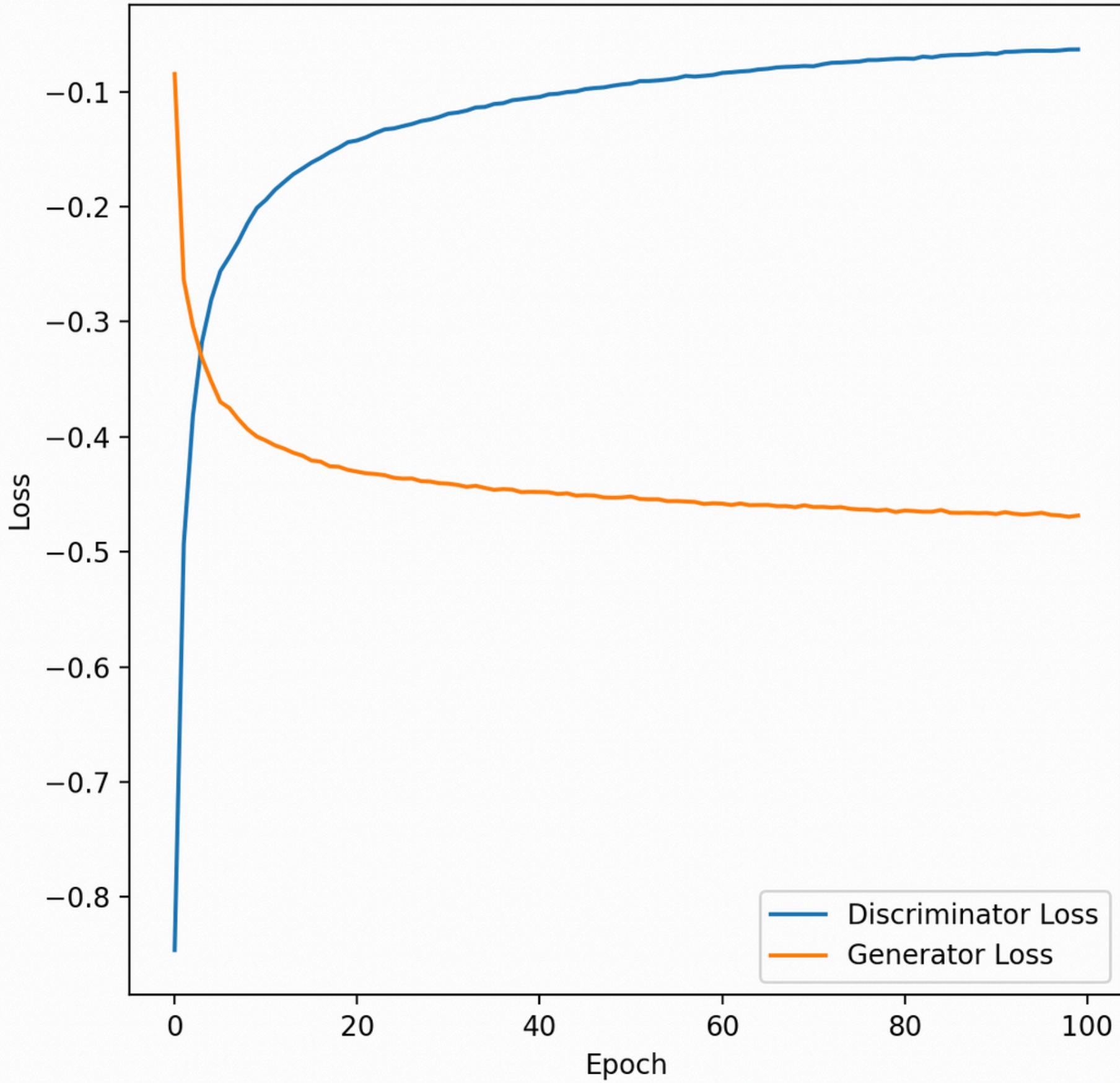
On optimise les données directement dans l'espace latent, gros risque de collapse et de réduire le recall, mais amélioration de la précision



- 
- 01 RÉSULTATS SATISFAISANTS
- 02 OPTIMISER LES HYPERPARAMÈTRES/ AMÉLIORER LE CALCUL DE LA NORME
- 03 ARBITRAGE ENTRE PRECISION ET RECALL
- 04 IL RESTE PLUSIEURS ARTEFACTS
- 05 APPLIQUER UNE MEILLEURE NORMALISATION DES VECTEURS LATENTS
- 06 ESSAYER SUR D'AUTRES TYPES DE GAN
- 07 MÉTHODE RAPIDE, POST TRAIN, IDÉALE POUR VISUALISER LES CHANGEMENTS

**MERCI!**

### Generator and Discriminator Loss



### Evolution of K\_eff

