WWW.DAUPHINE.PSL.EU

6/11/2024

GAN on MNIST

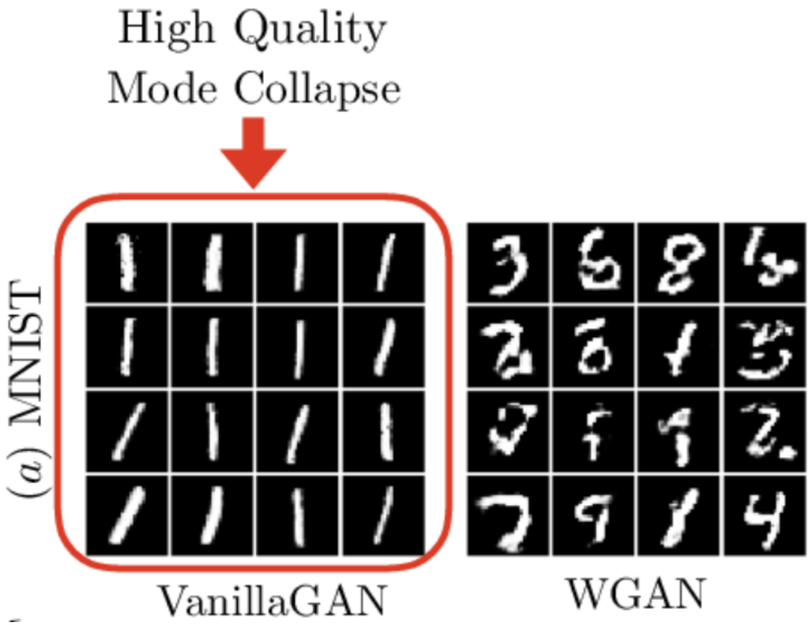
GANgineers

Matthieu NEAU
Efstathios CHATZILOZOS
Abderrahmane KASMI

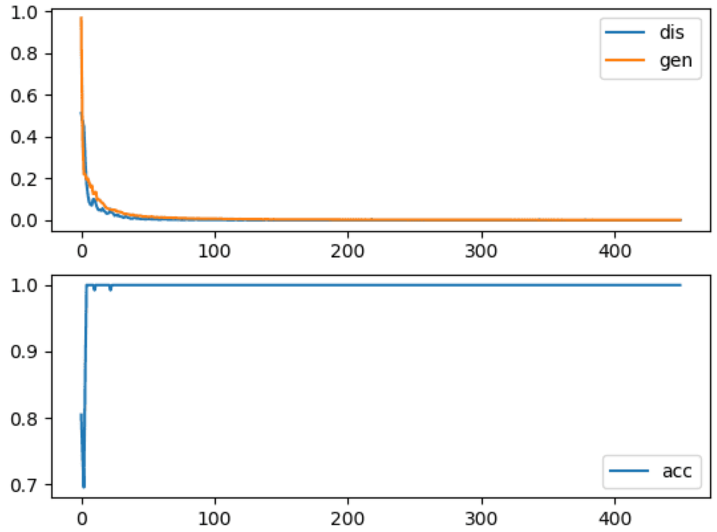
Dauphine | PSL
UNIVERSITÉ PARIS

What are Wasserstein GANs?

Solve 2 problems in Vanilla GANs:



Mode Collapse



Non convergence

source

source

WGANs innovations

New distance: EM distance

- Easier convergence of probability distributions
- Works on low dimensional manifolds.

$$D_{JS}(p\|q) = \frac{1}{2}D_{KL}(p\|\frac{p+q}{2}) + \frac{1}{2}D_{KL}(q\|\frac{p+q}{2})$$

$$\Rightarrow$$

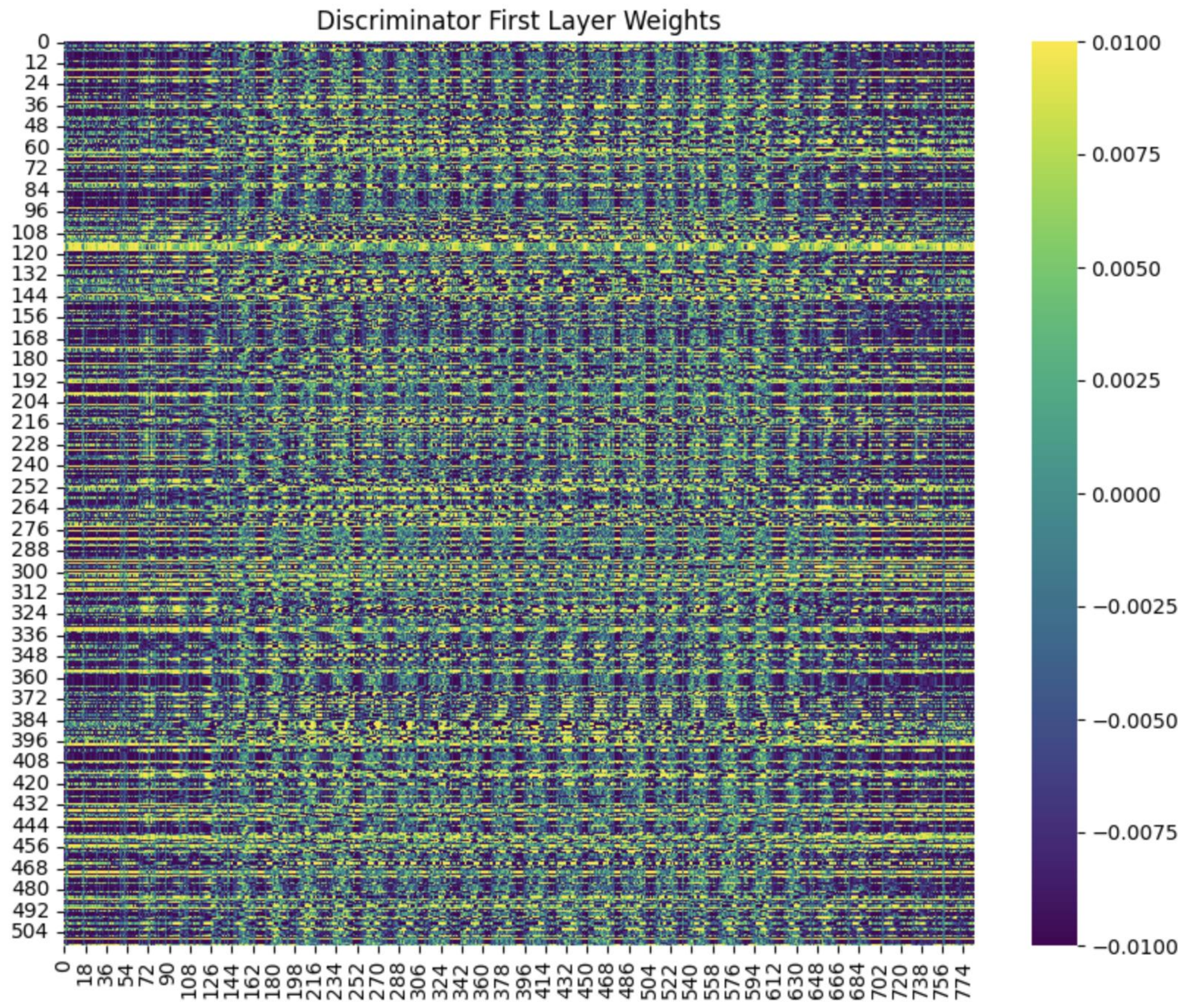
$$W(p_r, p_g) = \inf_{\gamma \sim \Pi(p_r, p_g)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|]$$

Dual problem:

$$W(p_r, p_g) = \frac{1}{K} \sup_{\|f\|_L \leq K} \mathbb{E}_{x\sim p_r}[f(x)] - \mathbb{E}_{x\sim p_g}[f(x)]$$

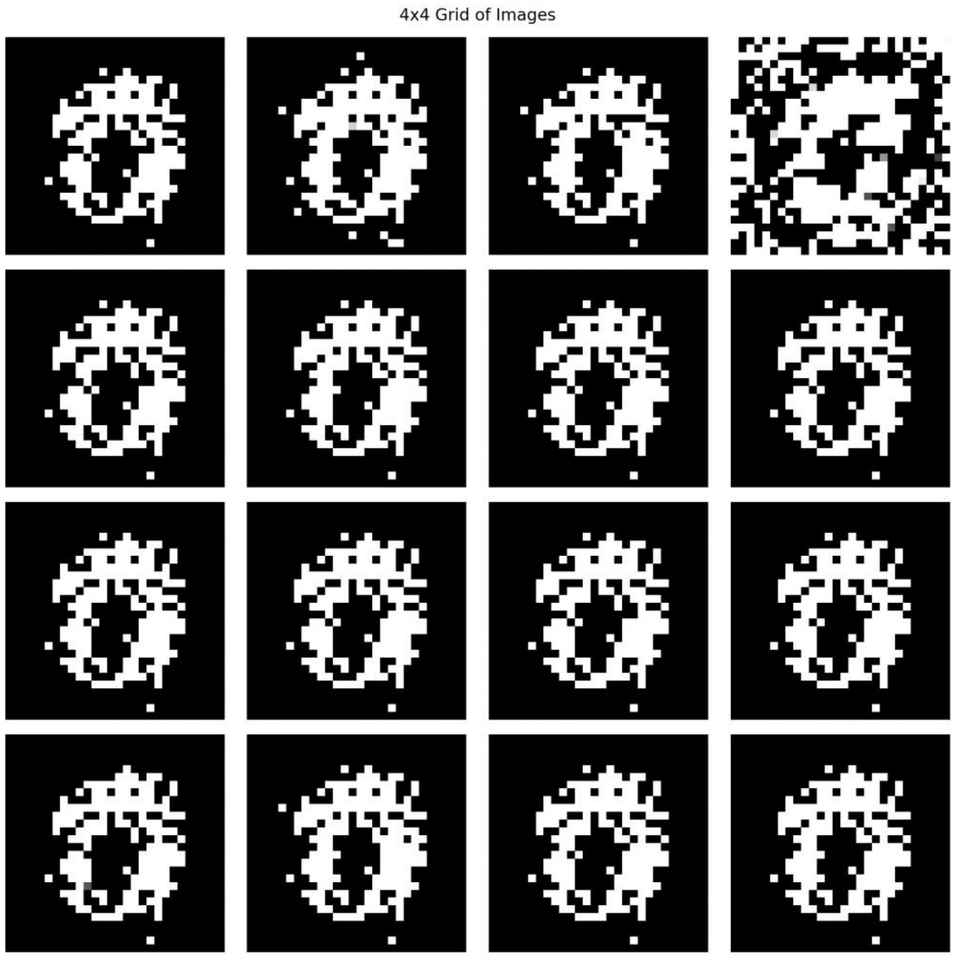
In practice, to ensure K Lipschitzianity, we **clip the discriminator's weights**

Adjusting weight clipping



- Clipping too small => vanishing gradients
- Clipping too high => Lipschitzianity not well enforced => Poor estimate of EM distance

Mode collapse on MNIST



- Easy to detect
- Solved by WGANs

WGAN-GP (Gradient Penalty)

WGANs require the critic to be Lipschitz continuous ($L \leq 1$).

- In the **original WGAN approach** it is enforced by **weight clipping**. However, it reduces the critic's capacity, leading to poor convergence and artifacts in generated samples.
- **Gradient Penalty** penalizes the gradient norm of the discriminator's output with respect to its input, encouraging it to be close to 1.

$$\hat{x} = \alpha \cdot x_{\text{real}} + (1 - \alpha) \cdot x_{\text{fake}}$$

$$\text{gradient\_penalty} = \lambda \cdot \mathbb{E}\left[(\|\nabla D(\hat{x})\|_2 - 1)^2\right]$$

$$D_{\text{loss}} = D_{\text{real}} - D_{\text{fake}} + \text{gradient\_penalty}$$

  - Maintains the full capacity of the critic (more accurate "real" vs "fake" distinction).
  - Provides smoother and more stable training process.
  - Improves convergence and quality of samples.

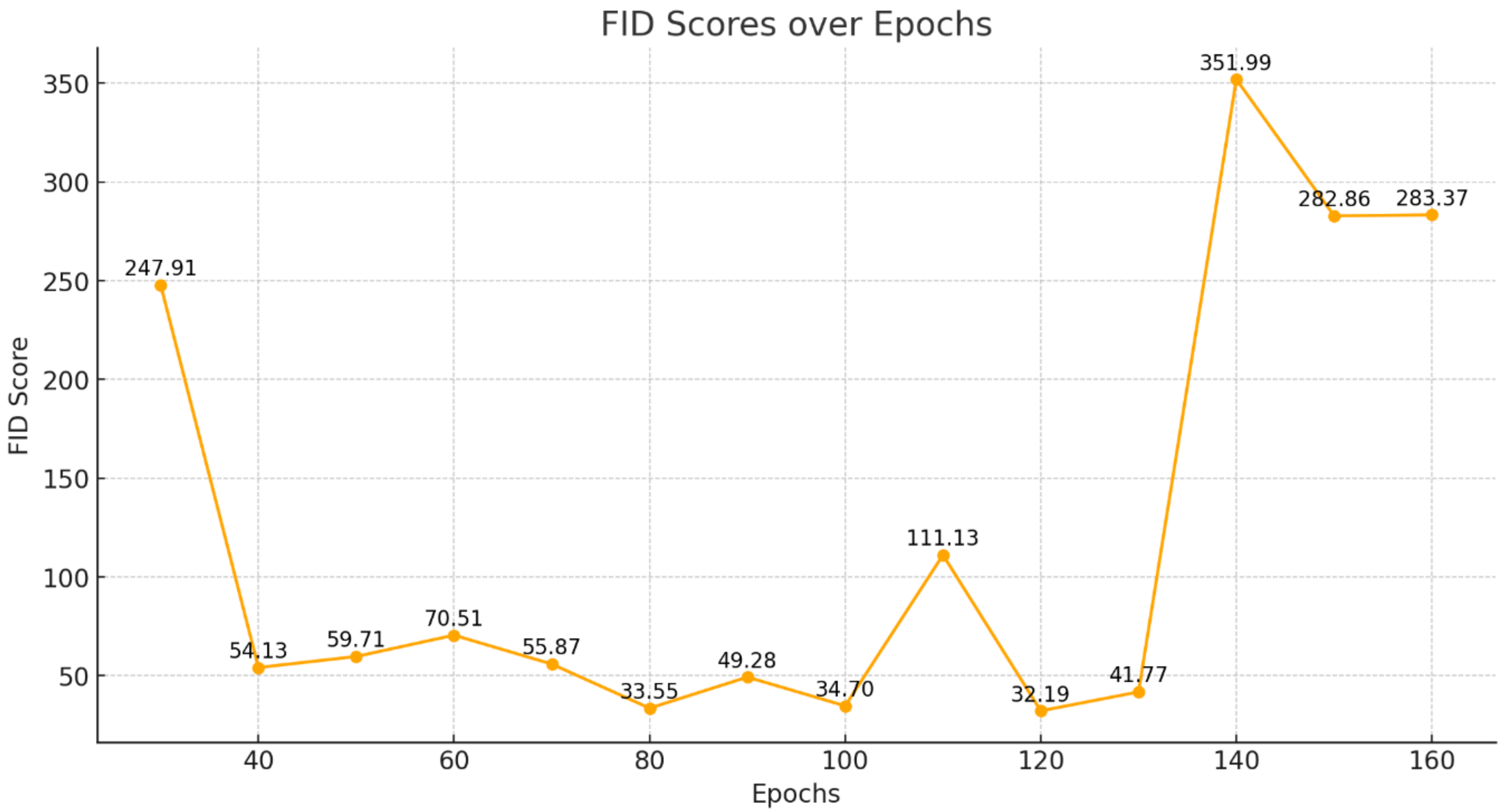
Discriminator Rejection Sampling (DRS)

- Post-processing technique applied to the generator's output to improve the quality of the generated samples by taking advantage of the discriminator.

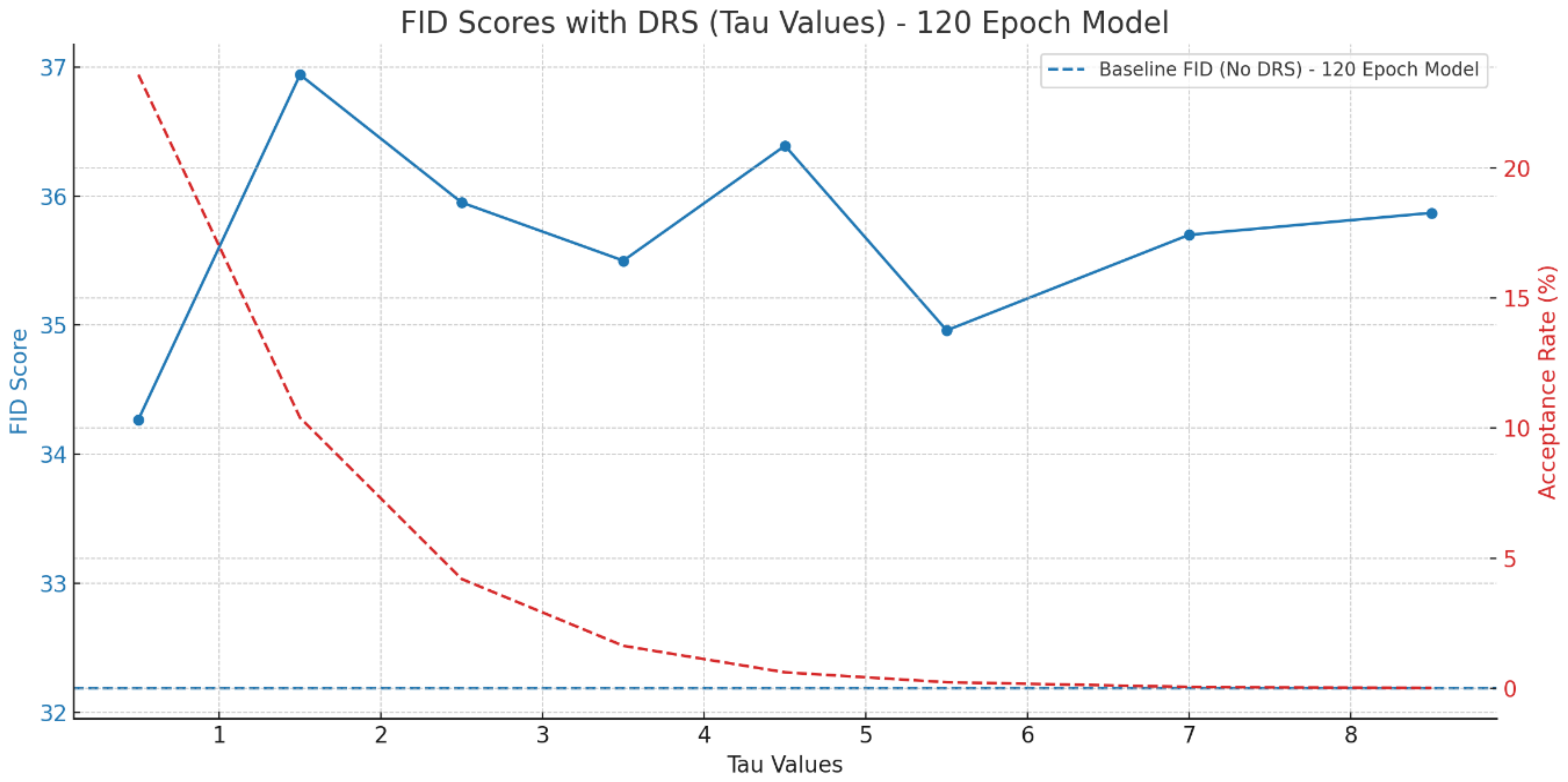
- Selectively accepts or rejects generated samples based on the discriminator's confidence, retaining more "realistic" samples according to the discriminator.

- Basic Steps:
  1. Generate a batch of fake samples using the Generator.
  2. Pass them through the Discriminator to obtain logits (confidence scores).
  3. Compute acceptance probability (e.g. sigmoid with a threshold τ).
  4. Sample from a Bernoulli distribution to select which ones to accept
  5. Repeat till the desired number of samples is obtained

WGAN-GP FID Results



DRS Effect on FID for WGAN-GP



Discriminator **Optimal** Transport [1]

$$W(p,q) = \max_{\|\tilde{D}\|_{\mathrm{Lip}} \leq 1} \left(\mathbb{E}_{x \sim p}\left[\tilde{D}(x)\right] - \mathbb{E}_{y \sim q}\left[\tilde{D}(y)\right] \right)$$

Theorem 1: *D* satisfies the following:*

$$\|\mathbf{D}^*\|_{\mathrm{Lip}} = 1$$

$$T(y) = \arg\min_x \left\{ \|x - y\|^2 - D^*(x) \right\}$$

Theorem 2: *Each objective function of GAN using logistic, or hinge, or identity loss with gradient penalty, provides lower bound of the mean discrepancy of $\tilde{D} = D/K$ between p and p_G*

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

Target space DOT

$$T^{D}_{\text{eff}}(y) = \arg\min_{x} \left\{ \|x - y\|^2 - \frac{1}{K_{\text{eff}}} D(x) \right\}$$

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

Algorithm 1 Target space optimal transport by gradient descent

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

Initialize $\boldsymbol{x} \leftarrow \boldsymbol{y}$

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

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

end for

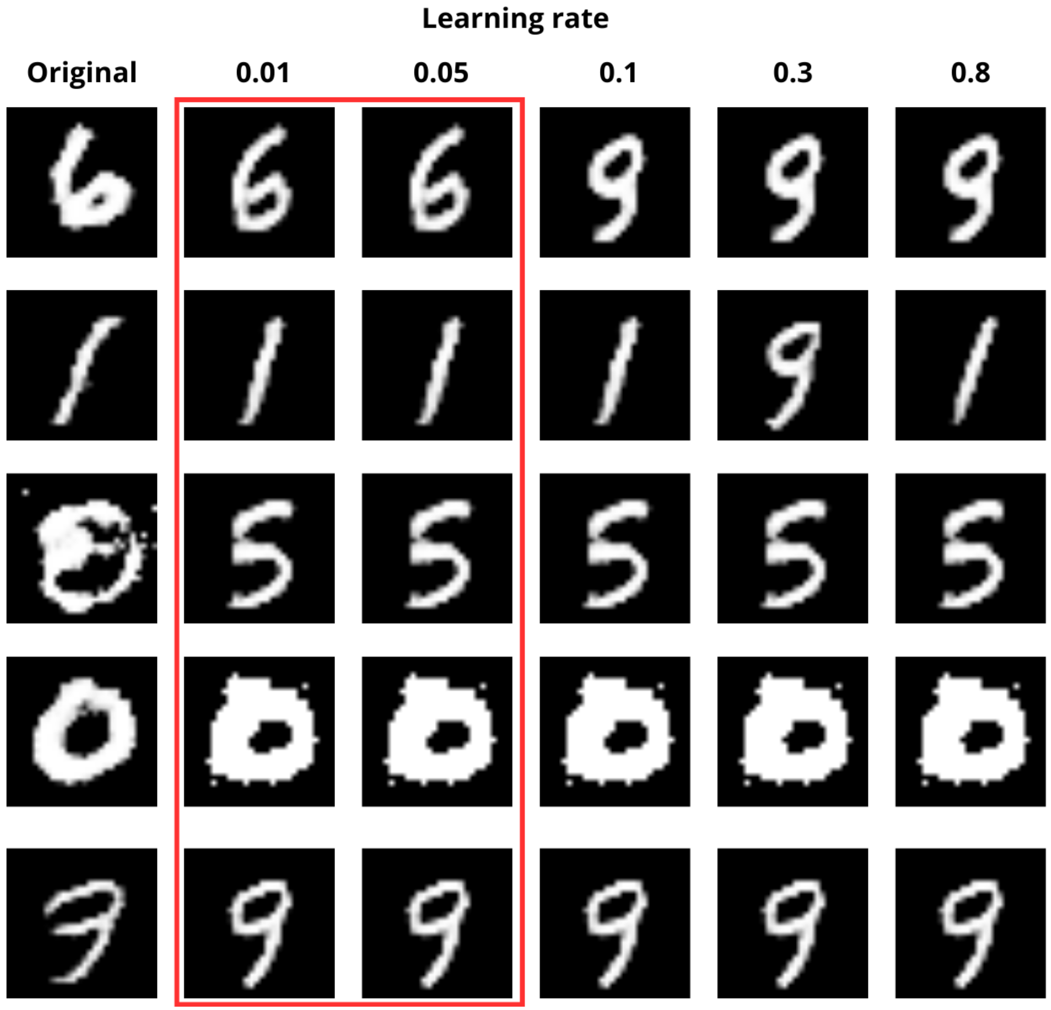
return $\boldsymbol{x}$

Latent space DOT

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

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

Latent space DOT - Result



With:
Lr = 0.01
Epochs = 10000

Vanilla GAN FID: 38.85
V GAN + DOT FID : 91.56

THANKS FOR LISTENING

Time for your questions

UNIVERSITÉ PARIS DAUPHINE - PSL
Place du Maréchal de Lattre de Tassigny – 75775 Paris cedex 16

References

[1] Tanaka, Akinori. "Discriminator optimal transport." Advances in Neural Information Processing Systems 32 (2019).

[2] Gulrajani, Ahmed, Arjovsky, Dumoulin, and Courville. "Improved Training of Wasserstein GANs." arXiv preprint arXiv:1704.00028 (2017).

[3] Azadi, Olsson, Darrell, Goodfellow, and Odena"Discriminator Rejection Sampling." arXiv preprint arXiv:1810.06758 (2019).