Deep Generative Models

Lecture 8

Roman Isachenko

AI Masters

2024, Spring

Recap of previous lecture

Likelihood-free learning

- Likelihood is not a perfect quality measure for generative model.
- Likelihood could be intractable.

Imagine we have two sets of samples

- $\mathcal{S}_1 = \{\mathbf{x}_i\}_{i=1}^{n_1} \sim \pi(\mathbf{x})$ – real samples;
- $\mathcal{S}_2 = \{\mathbf{x}_i\}_{i=1}^{n_2} \sim p(\mathbf{x}|\boldsymbol{\theta})$ – generated (or fake) samples.

Let define discriminative model (classifier):

$$p(y = 1|\mathbf{x}) = P\big(\{\mathbf{x} \sim \pi(\mathbf{x})\}\big); \quad p(y = 0|\mathbf{x}) = P\big(\{\mathbf{x} \sim p(\mathbf{x}|\boldsymbol{\theta})\}\big)$$

Assumption

Generative distribution $p(\mathbf{x}|\boldsymbol{\theta})$ equals to the true distribution $\pi(\mathbf{x})$ if we can not distinguish them using discriminative model $p(y|\mathbf{x})$. It means that $p(y = 1|\mathbf{x}) = 0.5$ for each sample $\mathbf{x}$.

Recap of previous lecture

- **Generator:** generative model $\mathbf{x} = \mathbf{G}(\mathbf{z})$, which makes generated sample more realistic.
- **Discriminator:** a classifier $D(\mathbf{x}) \in [0, 1]$, which distinguishes real samples from generated samples.

GAN optimality theorem

The minimax game

$$\min_{G} \max_{D} \Bigl[\underbrace{\mathbb{E}_{\pi(\mathbf{x})} \log D(\mathbf{x}) + \mathbb{E}_{p(\mathbf{z})} \log(1 - D(\mathbf{G}(\mathbf{z})))}_{V(G,D)} \Bigr]$$

has the global optimum $\pi(\mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta})$, in this case $D^*(\mathbf{x}) = 0.5$.

$$\min_{G} V(G, D^*) = \min_{G} \left[2JSD(\pi||p) - \log 4\right] = -\log 4, \quad \pi(\mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta}).$$

If the generator could be **any** function and the discriminator is **optimal** at every step, then the generator is **guaranteed to converge** to the data distribution.

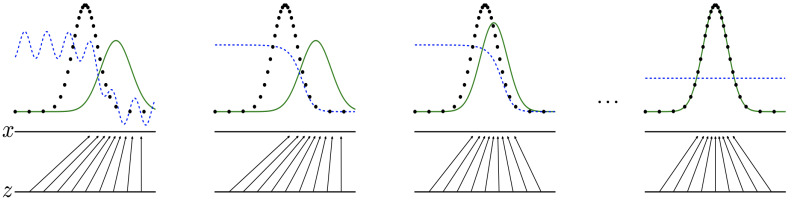
Goodfellow I. J. et al. Generative Adversarial Networks, 2014

Recap of previous lecture

- Generator updates are made in parameter space, discriminator is not optimal at every step.
- Generator and discriminator loss keeps oscillating during GAN training.

Objective

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\phi}} \left[\mathbb{E}_{\pi(\mathbf{x})} \log D_{\boldsymbol{\phi}}(\mathbf{x}) + \mathbb{E}_{p(\mathbf{z})} \log(1 - D_{\boldsymbol{\phi}}(\mathbf{G}_{\boldsymbol{\theta}}(\mathbf{z})))\right]$$



(a) (b) (c) (d)

Goodfellow I. J. et al. Generative Adversarial Networks, 2014

Recap of previous lecture

Main problems of standard GAN

- Vanishing gradients (solution: non-saturating GAN);
- Mode collapse (caused by Jensen-Shannon divergence).

Standard GAN

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\phi}} \left[\mathbb{E}_{\pi(\mathbf{x})} \log D_{\boldsymbol{\phi}}(\mathbf{x}) + \mathbb{E}_{p(\mathbf{z})} \log(1 - D_{\boldsymbol{\phi}}(\mathbf{G}_{\boldsymbol{\theta}}(\mathbf{z})))\right]$$

Informal theoretical results

The real images distribution $\pi(\mathbf{x})$ and the generated images distribution $p(\mathbf{x}|\boldsymbol{\theta})$ are low-dimensional and have disjoint supports. In this case

$$KL(\pi||p) = KL(p||\pi) = \infty, \quad JSD(\pi||p) = \log 2.$$

Goodfellow I. J. et al. Generative Adversarial Networks, 2014
Arjovsky M., Bottou L. Towards Principled Methods for Training Generative Adversarial Networks, 2017

Recap of previous lecture

Wasserstein distance

$$W(\pi, p) = \inf_{\gamma \in \Gamma(\pi, p)} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \gamma} \|\mathbf{x} - \mathbf{y}\| = \inf_{\gamma \in \Gamma(\pi, p)} \int \|\mathbf{x} - \mathbf{y}\| \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}$$

- $\gamma(\mathbf{x}, \mathbf{y})$ – transportation plan (the amount of "dirt" that should be transported from point $\mathbf{x}$ to point $\mathbf{y}$).
- $\Gamma(\pi, p)$ – the set of all joint distributions $\gamma(\mathbf{x}, \mathbf{y})$ with marginals π and p ($\int \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{x} = p(\mathbf{y})$, $\int \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{y} = \pi(\mathbf{x})$).
- $\gamma(\mathbf{x}, \mathbf{y})$ – the amount, $\|\mathbf{x} - \mathbf{y}\|$– the distance.

Theorem (Kantorovich-Rubinstein duality)

$$W(\pi || p) = \frac{1}{K} \max_{\|f\|_L \leq K} \left[\mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} f(\mathbf{x}) \right],$$

where $\|f\|_L \leq K$ are $K-$Lipschitz continuous functions ($f : \mathcal{X} \to \mathbb{R}$).

Arjovsky M., Chintala S., Bottou L. Wasserstein GAN, 2017

Outline

1. Lipschitzness of Wasserstein GAN critic
   - Wasserstein GAN
   - WGAN with Gradient Penalty
2. f-divergence minimization
3. Evaluation of likelihood-free models
   - Frechet Inception Distance (FID)

Outline

Wasserstein GAN

Theorem (Kantorovich-Rubinstein duality)

$$W(\pi || p) = \frac{1}{K} \max_{\|f\|_L \leq K} \left[\mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} f(\mathbf{x}) \right],$$

- Now we have to ensure that f is K-Lipschitz continuous.
- Let $f_\phi(\mathbf{x})$ be a feedforward neural network parametrized by ϕ.
- If parameters ϕ lie in a compact set $\boldsymbol{\Phi}$ then $f_\phi(\mathbf{x})$ will be K-Lipschitz continuous function.
- Let the parameters be clamped to a fixed box $\boldsymbol{\Phi} \in [-c, c]^d$ (e.x. $c = 0.01$) after each gradient update.

$$K \cdot W(\pi || p) = \max_{\|f\|_L \leq K} \left[\mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} f(\mathbf{x}) \right] \geq$$

$$\geq \max_{\phi \in \boldsymbol{\Phi}} \left[\mathbb{E}_{\pi(\mathbf{x})} f_\phi(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} f_\phi(\mathbf{x}) \right]$$

Arjovsky M., Chintala S., Bottou L. Wasserstein GAN, 2017

Wasserstein GAN

Standard GAN objective

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\phi}} \mathbb{E}_{\pi(\mathbf{x})} \log D_{\boldsymbol{\phi}}(\mathbf{x}) + \mathbb{E}_{p(\mathbf{z})} \log(1 - D_{\boldsymbol{\phi}}(\mathbf{G}_{\boldsymbol{\theta}}(\mathbf{z})))$$

WGAN objective

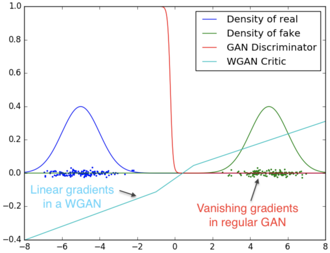
$$\min_{\boldsymbol{\theta}} W(\pi || p) \approx \min_{\boldsymbol{\theta}} \max_{\boldsymbol{\phi} \in \boldsymbol{\Phi}} \left[\mathbb{E}_{\pi(\mathbf{x})} f_{\boldsymbol{\phi}}(\mathbf{x}) - \mathbb{E}_{p(\mathbf{z})} f_{\boldsymbol{\phi}}(\mathbf{G}_{\boldsymbol{\theta}}(\mathbf{z})) \right].$$

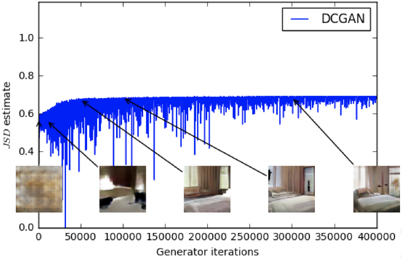
- Discriminator D is similar to the function f, but not the same (it is not a classifier anymore). In the WGAN model, function f is usually called **critic**.
- *"Weight clipping is a clearly terrible way to enforce a Lipschitz constraint"*. If the clipping parameter c is too large, it is hard to train the critic till optimality. If the clipping parameter c is too small, it could lead to vanishing gradients.

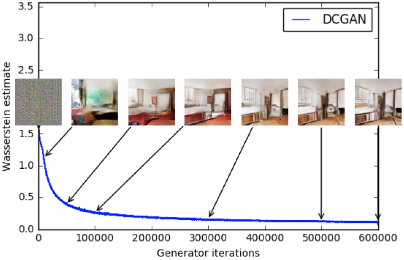
Arjovsky M., Chintala S., Bottou L. Wasserstein GAN, 2017

Wasserstein GAN

- WGAN has non-zero gradients for disjoint supports.
- $JSD(\pi||p)$ correlates poorly with the sample quality. Stays constast nearly maximum value $\log 2 \approx 0.69$.
- $W(\pi||p)$ is highly correlated with the sample quality.



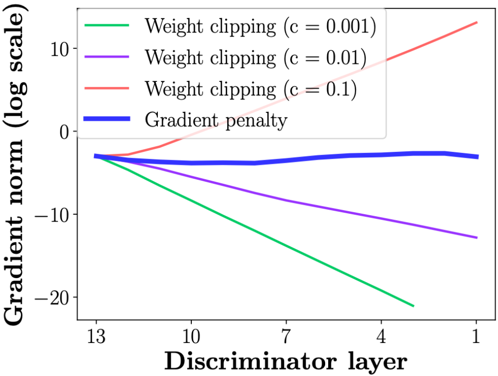


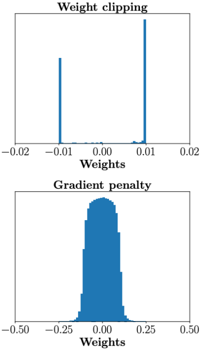


Arjovsky M., Chintala S., Bottou L. Wasserstein GAN, 2017

Outline

Wasserstein GAN with Gradient Penalty





Weight clipping analysis

- The gradients either grow or decay exponentially.
- Gradient penalty makes the gradients more stable.

Gulrajani I. et al. Improved Training of Wasserstein GANs, 2017

Wasserstein GAN with Gradient Penalty

Theorem

Let $\pi(\mathbf{x})$ and $p(\mathbf{x})$ be two distribution in $\mathcal{X}$, a compact metric space. Let γ be the optimal transportation plan between $\pi(\mathbf{x})$ and $p(\mathbf{x})$. Then

1. there is 1-Lipschitz function f^* which is the optimal solution of
$$\max_{\|f\|_L \leq 1} \left[\mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} f(\mathbf{x}) \right].$$
2. if f^* is differentiable, $\gamma(\mathbf{y} = \mathbf{z}) = 0$ and $\hat{\mathbf{x}}_t = t\mathbf{y} + (1-t)\mathbf{z}$ with $\mathbf{y} \sim \pi(\mathbf{x})$, $\mathbf{z} \sim p(\mathbf{x}|\boldsymbol{\theta})$, $t \in [0, 1]$ it holds that
$$\mathbb{P}_{(\mathbf{y},\mathbf{z})\sim\gamma} \left[\nabla f^*(\hat{\mathbf{x}}_t) = \frac{\mathbf{z} - \hat{\mathbf{x}}_t}{\|\mathbf{z} - \hat{\mathbf{x}}_t\|} \right] = 1.$$

Corollary

f^* has gradient norm 1 almost everywhere under $\pi(\mathbf{x})$ and $p(\mathbf{x})$.

Gulrajani I. et al. Improved Training of Wasserstein GANs, 2017

Wasserstein GAN with Gradient Penalty

A differentiable function is 1-Lipschtiz if and only if it has gradients with norm at most 1 everywhere.

Gradient penalty

$$W(\pi||p) = \underbrace{\mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} f(\mathbf{x})}_{\text{original critic loss}} + \lambda \underbrace{\mathbb{E}_{U[0,1]} \left[(\|\nabla f(\hat{\mathbf{x}})\|_2 - 1)^2 \right]}_{\text{gradient penalty}},$$

- Samples $\hat{\mathbf{x}}_t = t \cdot \mathbf{y} + (1 - t) \cdot \mathbf{z}$ with $t \in [0, 1]$ are uniformly sampled along straight lines between pairs of points: $\mathbf{y}$ from the data distribution $\pi(\mathbf{x})$ and $\mathbf{z}$ from the generator distribution $p(\mathbf{x}|\boldsymbol{\theta})$.
- Enforcing the unit gradient norm constraint everywhere is intractable, it turns out to be sifficient to enforce it only along these straight lines.

Gulrajani I. et al. Improved Training of Wasserstein GANs, 2017

Outline

1. Lipschitzness of Wasserstein GAN critic
   Wasserstein GAN
   WGAN with Gradient Penalty

2. f-divergence minimization

3. Evaluation of likelihood-free models
   Frechet Inception Distance (FID)

Outline

Divergences

- Forward KL divergence in maximum likelihood estimation.
- Reverse KL in variational inference.
- JS divergence in standard GAN.
- Wasserstein distance in WGAN.

What is a divergence?

Let $\mathcal{P}$ be the set of all possible probability distributions. Then $D : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ is a divergence if

- $D(\pi || p) \geq 0$ for all $\pi, p \in \mathcal{P}$;
- $D(\pi || p) = 0$ if and only if $\pi \equiv p$.

General divergence minimization task

$$\min_p D(\pi || p)$$

Chalenge

We do not know the real distribution $\pi(\mathbf{x})$!

f-divergence family

f-divergence

$$D_f(\pi||p) = \mathbb{E}_{p(\mathbf{x})} f\left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})}\right) = \int p(\mathbf{x}) f\left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x}.$$

Here $f : \mathbb{R}_+ \to \mathbb{R}$ is a convex, lower semicontinuous function satisfying $f(1) = 0$.

Name	$D_f(P\|Q)$	Generator $f(u)$
Kullback-Leibler	$\int p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x$	$u \log u$
Reverse KL	$\int q(x) \log \frac{q(x)}{p(x)} \, \mathrm{d}x$	$-\log u$
Pearson χ^2	$\int \frac{(q(x)-p(x))^2}{p(x)} \, \mathrm{d}x$	$(u-1)^2$
Squared Hellinger	$\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}x$	$(\sqrt{u}-1)^2$
Jensen-Shannon	$\frac{1}{2} \int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} \, \mathrm{d}x$	$-(u+1) \log \frac{1+u}{2} + u \log u$
GAN	$\int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} \, \mathrm{d}x - \log(4)$	$u \log u - (u+1) \log(u+1)$

Nowozin S., Cseke B., Tomioka R. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization, 2016

f-divergence family

Fenchel conjugate

$$f^*(t) = \sup_{u \in \mathrm{dom}_f} \left(ut - f(u) \right), \quad f(u) = \sup_{t \in \mathrm{dom}_{f^*}} \left(ut - f^*(t) \right)$$

Important property: $f^{**} = f$ for convex f.

f-divergence

$$D_f(\pi || p) = \mathbb{E}_{p(\mathbf{x})} f\left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})}\right) = \int p(\mathbf{x}) f\left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x} =$$

$$= \int p(\mathbf{x}) \sup_{t \in \mathrm{dom}_{f^*}} \left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})} t - f^*(t) \right) d\mathbf{x} =$$

$$= \int \sup_{t \in \mathrm{dom}_{f^*}} \left(\pi(\mathbf{x}) t - p(\mathbf{x}) f^*(t) \right) d\mathbf{x}.$$

Here we seek value of t, which gives us maximum value of $\pi(\mathbf{x})t - p(\mathbf{x})f^*(t)$, for each data point $\mathbf{x}$.

Nowozin S., Cseke B., Tomioka R. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization, 2016

f-divergence family

f-divergence

$$D_f(\pi||p) = \mathbb{E}_{p(\mathbf{x})} f\left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})}\right) = \int p(\mathbf{x}) f\left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x}.$$

Variational f-divergence estimation

$$\begin{aligned} D_f(\pi||p) = \int \sup_{t \in \mathrm{dom}_{f^*}} \left(\pi(\mathbf{x}) t - p(\mathbf{x}) f^*(t)\right) d\mathbf{x} \geq \\ \geq \sup_{T \in \mathcal{T}} \int \left(\pi(\mathbf{x}) T(\mathbf{x}) - p(\mathbf{x}) f^*(T(\mathbf{x}))\right) d\mathbf{x} = \\ = \sup_{T \in \mathcal{T}} \left[\mathbb{E}_\pi T(\mathbf{x}) - \mathbb{E}_p f^*(T(\mathbf{x}))\right] \end{aligned}$$

This is a lower bound because of Jensen inequality and restricted class of functions $\mathcal{T} : \mathcal{X} \to \mathbb{R}$.

Nowozin S., Cseke B., Tomioka R. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization, 2016

f-divergence family

Variational divergence estimation

$$D_f(\pi||p) \geq \sup_{T \in \mathcal{T}} \left[\mathbb{E}_\pi T(\mathbf{x}) - \mathbb{E}_p f^*(T(\mathbf{x}))\right]$$

The lower bound is tight for $T^*(\mathbf{x}) = f'\left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})}\right)$.

Example (JSD)

- Let define function f and its conjugate f^*

$$f(u) = u \log u - (u+1)\log(u+1), \quad f^*(t) = -\log(1-e^t).$$

- Let reparametrize $T(\mathbf{x}) = \log D(\mathbf{x})$.

$$\min_G \max_D \left[\mathbb{E}_{\pi(\mathbf{x})} \log D(\mathbf{x}) + \mathbb{E}_{p(\mathbf{z})} \log(1 - D(\mathbf{G}(\mathbf{z})))\right]$$

Nowozin S., Cseke B., Tomioka R. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization, 2016

f-divergence family

Variational divergence estimation

$$D_f(\pi||p) \geq \sup_{T \in \mathcal{T}} \left[\mathbb{E}_\pi T(\mathbf{x}) - \mathbb{E}_p f^*(T(\mathbf{x}))\right]$$

Note: To evaluate lower bound we only need samples from $\pi(\mathbf{x})$ and $p(\mathbf{x})$. Hence, we could fit implicit generative model.

(a) GAN (b) KL (c) Squared Hellinger

Nowozin S., Cseke B., Tomioka R. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization, 2016

Outline

1. Lipschitzness of Wasserstein GAN critic
   - Wasserstein GAN
   - WGAN with Gradient Penalty
2. f-divergence minimization
3. Evaluation of likelihood-free models
   - Frechet Inception Distance (FID)

Evaluation of likelihood-free models

How to evaluate generative models?

Likelihood-based models

- Split data to train/val/test.
- Fit model on the train part.
- Tune hyperparameters on the validation part.
- Evaluate generalization by reporting likelihoods on the test set.

Not all models have tractable likelihoods

- VAE: compare ELBO values.
- GAN: ???

Evaluation of likelihood-free models

Let take some pretrained image classification model to get the conditional label distribution $p(y|\mathbf{x})$ (e.g. ImageNet classifier).

What do we want from samples?

- **Sharpness**

Low sharpness

High sharpness

The conditional distribution $p(y|\mathbf{x})$ should have low entropy (each image $\mathbf{x}$ should have distinctly recognizable object).

- **Diversity**

Low diversity

High diversity

The marginal distribution $p(y) = \int p(y|\mathbf{x})p(\mathbf{x})d\mathbf{x}$ should have high entropy (there should be as many classes generated as possible).

---

image credit: https://deepgenerativemodels.github.io

Outline

1. Lipschitzness of Wasserstein GAN critic
   - Wasserstein GAN
   - WGAN with Gradient Penalty

2. f-divergence minimization

3. Evaluation of likelihood-free models
   - Frechet Inception Distance (FID)

Frechet Inception Distance (FID)

Wasserstein metric

$$W_s(\pi, p) = \inf_{\gamma \in \Gamma(\pi, p)} \left(\mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim \gamma} \|\mathbf{x} - \mathbf{y}\|^s \right)^{1/s}$$

Theorem

If $\pi(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\pi, \boldsymbol{\Sigma}_\pi)$, $p(\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$, then

$$W_2^2(\pi, p) = \|\boldsymbol{\mu}_\pi - \boldsymbol{\mu}_p\|_2^2 + \mathrm{tr}\left[\boldsymbol{\Sigma}_\pi + \boldsymbol{\Sigma}_p - 2 \left(\boldsymbol{\Sigma}_\pi^{1/2} \boldsymbol{\Sigma}_p \boldsymbol{\Sigma}_\pi^{1/2} \right)^{1/2} \right]$$

Frechet Inception Distance

$$\mathrm{FID}(\pi, p) = \|\boldsymbol{\mu}_\pi - \boldsymbol{\mu}_p\|_2^2 + \mathrm{tr}\left[\boldsymbol{\Sigma}_\pi + \boldsymbol{\Sigma}_p - 2 \left(\boldsymbol{\Sigma}_\pi^{1/2} \boldsymbol{\Sigma}_p \boldsymbol{\Sigma}_\pi^{1/2} \right)^{1/2} \right]$$

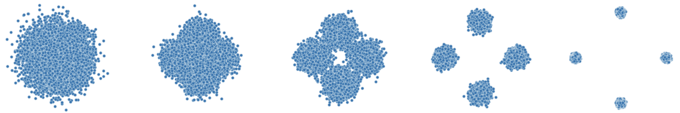
Representations are the outputs of the intermediate layer from the pretrained classification model.

Heusel M. et al. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, 2017

Frechet Inception Distance (FID)

$$\mathrm{FID}(\pi, p) = \|\boldsymbol{\mu}_\pi - \boldsymbol{\mu}_p\|_2^2 + \mathrm{tr}\left[\boldsymbol{\Sigma}_\pi + \boldsymbol{\Sigma}_p - 2\left(\boldsymbol{\Sigma}_\pi^{1/2}\boldsymbol{\Sigma}_p\boldsymbol{\Sigma}_\pi^{1/2}\right)^{1/2}\right]$$

- Needs a large sample size for evaluation.
- Calculation of FID is slow.
- High dependence on the pretrained classification model.
- Uses the normality assumption!



FD	0.0	0.0	0.0	0.0	0.0

Jayasumana S. et al. Rethinking FID: Towards a Better Evaluation Metric for Image Generation, 2024

Summary

- Wasserstein GAN uses Kantorovich-Rubinstein duality for getting Earth Mover distance as model objective.
- Weight clipping is a terrible way to enforce Lipschitzness. Gradient Penalty adds regularizer to loss that uses neccessary conditions for optimal critic.
- f-divergence family is a unified framework for divergence minimization, which uses variational approximation. Standard GAN is a special case of it.
- We need measure of quality for implicit models like GANs. One way is to analyze sharpness and diversity of samples.
- Frechet Inception Distance is the most popular metric for GAN evaluation.