

Deep Generative Models

Lecture 9

Roman Isachenko

Moscow Institute of Physics and Technology

2023, Autumn

Recap of previous lecture

Main problems of standard GAN

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

Standard GAN

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

Informal theoretical results

The real images distribution $\pi(x)$ and the generated images distribution $p(x|\theta)$ 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)

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

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

Recap of previous lecture

WGAN objective

$$\min_{\theta} W(\pi||p) = \min_{\theta} \max_{\phi \in \Phi} [\mathbb{E}_{\pi(x)} f_{\phi}(x) - \mathbb{E}_{p(z)} f_{\phi}(G_{\theta}(z))].$$

- ▶ Function f in WGAN is usually called *critic*.
- ▶ If parameters ϕ lie in a compact set $\Phi \in [-c, c]^d$ then $f(x, \phi)$ will be K -Lipschitz continuous function.

$$\begin{aligned} K \cdot W(\pi||p) &= \max_{\|f\|_L \leq K} [\mathbb{E}_{\pi(x)} f(x) - \mathbb{E}_{p(x)} f(x)] \geq \\ &\geq \max_{\phi \in \Phi} [\mathbb{E}_{\pi(x)} f_{\phi}(x) - \mathbb{E}_{p(x)} f_{\phi}(x)] \end{aligned}$$

"Weight clipping is a clearly terrible way to enforce a Lipschitz constraint"

Recap of previous lecture

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

$$\mathbb{P}_{(\mathbf{y}, \mathbf{z}) \sim \gamma} \left[\nabla f^*(\hat{\mathbf{x}}_t) = \frac{\mathbf{z} - \hat{\mathbf{x}}_t}{\|\mathbf{z} - \hat{\mathbf{x}}_t\|} \right] = 1.$$

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 $\hat{\mathbf{x}}_t = t\mathbf{y} + (1 - t)\mathbf{z}$ 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 $p(\mathbf{x}|\theta)$.

Recap of previous lecture

$$f(\mathbf{x}, \phi) = \mathbf{W}_{K+1} \sigma_K (\mathbf{W}_K \sigma_{K-1} (\dots \sigma_1 (\mathbf{W}_1 \mathbf{x}) \dots)).$$

- ▶ σ_k is a pointwise nonlinearities. We assume that $\|\sigma_k\|_L = 1$ (it holds for ReLU).
- ▶ $\mathbf{g}(\mathbf{x}) = \mathbf{W}\mathbf{x}$ is a linear transformation ($\nabla \mathbf{g}(\mathbf{x}) = \mathbf{W}$).

$$\|\mathbf{g}\|_L = \sup_{\mathbf{x}} \|\nabla \mathbf{g}(\mathbf{x})\|_2 = \|\mathbf{W}\|_2.$$

Critic spectral norm

$$\|f\|_L \leq \|\mathbf{W}_{K+1}\|_2 \cdot \prod_{k=1}^K \|\sigma_k\|_L \cdot \|\mathbf{W}_k\|_2 = \prod_{k=1}^{K+1} \|\mathbf{W}_k\|_2.$$

Spectral Normalization GAN

If we replace the weights in the critic $f(\mathbf{x}, \phi)$ by $\mathbf{W}_k^{SN} = \mathbf{W}_k / \|\mathbf{W}_k\|_2$, we will get $\|f\|_L \leq 1$.

Power iteration approximates the value of $\|\mathbf{W}\|_2$.

Outline

1. f-divergence minimization
2. Evaluation of likelihood-free models
 - Inception score
 - Frechet Inception Distance
 - Precision-Recall

Outline

1. f-divergence minimization
2. Evaluation of likelihood-free models
 - Inception score
 - Frechet Inception Distance
 - Precision-Recall

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

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

General divergence minimization task

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

Challenge

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

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

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

f-divergence family

Fenchel conjugate

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

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

f-divergence

$$\begin{aligned} D_f(\pi || p) &= \mathbb{E}_{p(\mathbf{x})} f\left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})}\right) = \int p(\mathbf{x}) \color{purple}{f\left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})}\right)} d\mathbf{x} = \\ &= \int p(\mathbf{x}) \sup_{t \in \text{dom}_{f^*}} \left(\frac{\pi(\mathbf{x})}{p(\mathbf{x})} t - f^*(t) \right) d\mathbf{x} = \\ &= \int \sup_{t \in \text{dom}_{f^*}} (\pi(\mathbf{x})t - p(\mathbf{x})f^*(t)) d\mathbf{x}. \end{aligned}$$

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} .

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

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

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

f-divergence family

Variational divergence estimation

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

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

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

f-divergence family

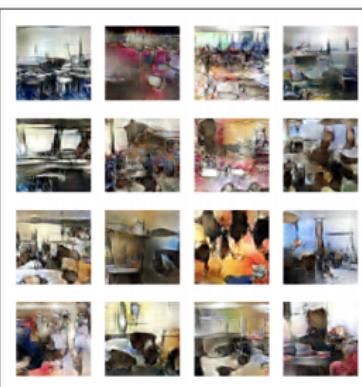
Variational divergence estimation

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

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

Outline

1. f-divergence minimization
2. Evaluation of likelihood-free models
 - Inception score
 - Frechet Inception Distance
 - Precision-Recall

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's take some pretrained image classification model to get the conditional label distribution $p(y|x)$ (e.g. ImageNet classifier).

What do we want from samples?

- ▶ Sharpness



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

- ▶ Diversity



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

Evaluation of likelihood-free models

What do we want from samples?

- ▶ **Sharpness.** The conditional distribution $p(y|x)$ should have low entropy (each image x should have distinctly recognizable object).
- ▶ **Diversity.** The marginal distribution $p(y) = \int p(y|x)p(x)dx$ should have high entropy (there should be as many classes generated as possible).

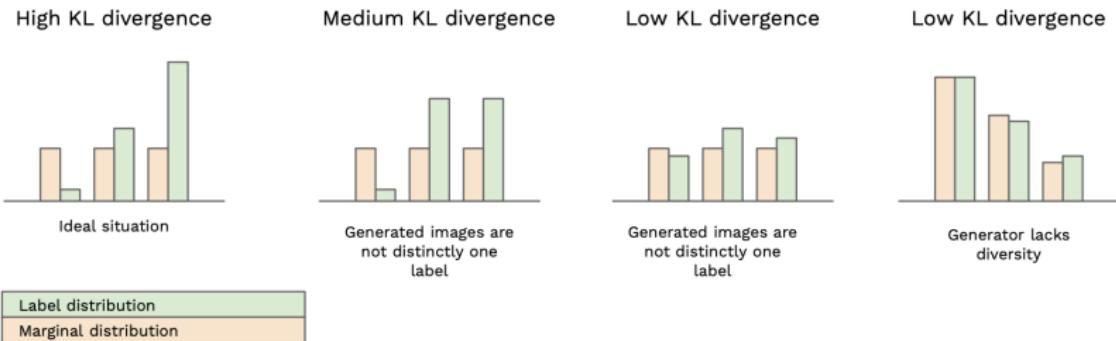


image credit: <https://medium.com/octavian-ai/a-simple-explanation-of-the-inception-score-372dff6a8c7a>

Outline

1. f-divergence minimization
2. Evaluation of likelihood-free models

Inception score

Frechet Inception Distance

Precision-Recall

Evaluation of likelihood-free models

What do we want from samples?

- ▶ Sharpness \Rightarrow low $H(y|\mathbf{x}) = - \sum_y \int_{\mathbf{x}} p(y, \mathbf{x}) \log p(y|\mathbf{x}) d\mathbf{x}$.
- ▶ Diversity \Rightarrow high $H(y) = - \sum_y p(y) \log p(y)$.

Inception Score

$$\begin{aligned} IS &= \exp(H(y) - H(y|\mathbf{x})) \\ &= \exp \left(- \sum_y p(y) \log p(y) + \sum_y \int_{\mathbf{x}} p(y, \mathbf{x}) \log p(y|\mathbf{x}) d\mathbf{x} \right) \\ &= \exp \left(\sum_y \int_{\mathbf{x}} p(y, \mathbf{x}) \log \frac{p(y|\mathbf{x})}{p(y)} d\mathbf{x} \right) \\ &= \exp \left(\mathbb{E}_{\mathbf{x}} \sum_y p(y|\mathbf{x}) \log \frac{p(y|\mathbf{x})}{p(y)} \right) = \exp (\mathbb{E}_{\mathbf{x}} KL(p(y|\mathbf{x}) || p(y))) \end{aligned}$$

Outline

1. f-divergence minimization
2. Evaluation of likelihood-free models

Inception score

Frechet Inception Distance

Precision-Recall

Evaluation of likelihood-free models

Theorem (informal)

If $\pi(\mathbf{x})$ and $p(\mathbf{x}|\theta)$ has moment generation functions then

$$\pi(\mathbf{x}) = p(\mathbf{x}|\theta) \Leftrightarrow \mathbb{E}_\pi \mathbf{x}^k = \mathbb{E}_p \mathbf{x}^k, \quad \forall k \geq 1.$$

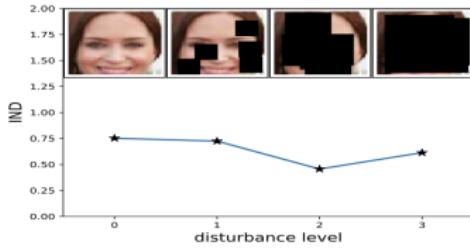
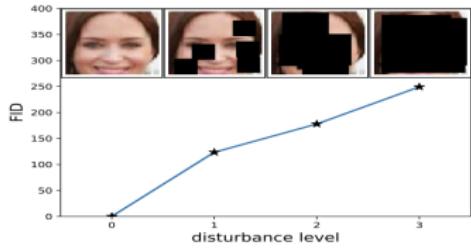
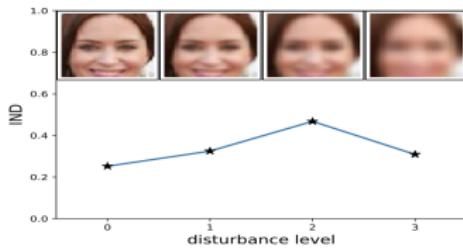
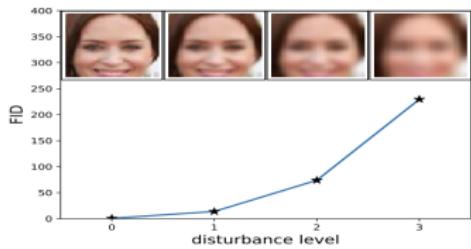
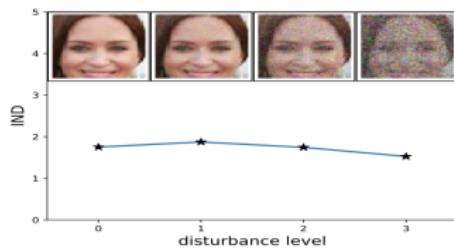
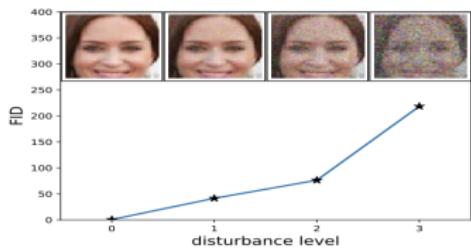
This is intractable to calculate all moments.

Frechet Inception Distance

$$FID(\pi, p) = \|\mathbf{m}_\pi - \mathbf{m}_p\|_2^2 + \text{Tr} \left(\boldsymbol{\Sigma}_\pi + \boldsymbol{\Sigma}_p - 2\sqrt{\boldsymbol{\Sigma}_\pi \boldsymbol{\Sigma}_p} \right)$$

- ▶ Representations are the outputs of the intermediate layer from the pretrained classification model.
- ▶ $\mathbf{m}_\pi, \boldsymbol{\Sigma}_\pi$ are the mean vector and the covariance matrix of feature representations for samples from $\pi(\mathbf{x})$
- ▶ $\mathbf{m}_p, \boldsymbol{\Sigma}_p$ are the mean vector and the covariance matrix of feature representations for samples from $p(\mathbf{x}|\theta)$.

Evaluation of likelihood-free models



Limitations

Inception Score

$$IS = \exp(\mathbb{E}_{\mathbf{x}} KL(p(y|\mathbf{x}) || p(y)))$$

- ▶ If generator produces images with a different set of labels from the classifier training set, IS will be low.
- ▶ If generator produces one image per class, the IS will be perfect (there is no measure of intra-class diversity).

Frechet Inception Distance

$$FID = \|\mathbf{m}_\pi - \mathbf{m}_p\|_2^2 + \text{Tr} \left(\boldsymbol{\Sigma}_\pi + \boldsymbol{\Sigma}_p - 2\sqrt{\boldsymbol{\Sigma}_\pi \boldsymbol{\Sigma}_p} \right)$$

- ▶ Needs a large sample size for evaluation.
- ▶ Calculation of FID is slow.
- ▶ Estimates only two sample moments.

Both scores depend on the pretrained classifier $p(y|\mathbf{x})$.

Barratt S., Sharma R. A Note on the Inception Score, 2018

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

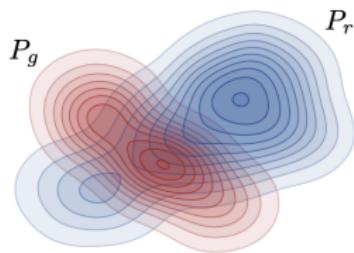
Outline

1. f-divergence minimization
2. Evaluation of likelihood-free models
 - Inception score
 - Frechet Inception Distance
 - Precision-Recall

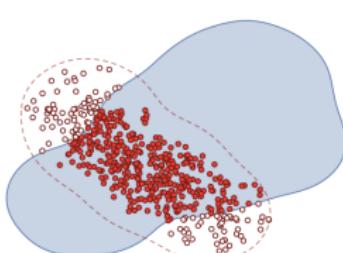
Precision-Recall for Generative Models

What do we want from samples

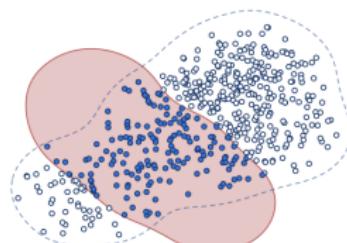
- ▶ **Sharpness:** generated samples should be of high quality.
- ▶ **Diversity:** their variation should match that observed in the training set.



(a) Example distributions



(b) Precision



(c) Recall

- ▶ **Precision** denotes the fraction of generated images that are realistic.
- ▶ **Recall** measures the fraction of the training data manifold covered by the generator.

Precision-Recall for generative models

- ▶ $\mathcal{S}_\pi = \{\mathbf{x}_i\}_{i=1}^n \sim \pi(\mathbf{x})$ – real samples;
- ▶ $\mathcal{S}_p = \{\mathbf{x}_i\}_{i=1}^n \sim p(\mathbf{x}|\theta)$ – generated samples.

Embed samples using pretrained classifier network (as previously):

$$\mathcal{G}_\pi = \{\mathbf{g}_i\}_{i=1}^n, \quad \mathcal{G}_p = \{\mathbf{g}_i\}_{i=1}^n.$$

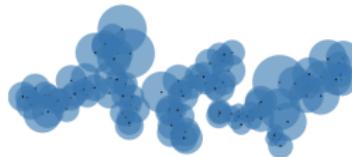
Define binary function:

$$f(\mathbf{g}, \mathcal{G}) = \begin{cases} 1, & \text{if exists } \mathbf{g}' \in \mathcal{G} : \|\mathbf{g} - \mathbf{g}'\|_2 \leq \|\mathbf{g}' - \text{NN}_k(\mathbf{g}', \mathcal{G})\|_2; \\ 0, & \text{otherwise.} \end{cases}$$

$$\text{Precision}(\mathcal{G}_\pi, \mathcal{G}_p) = \frac{1}{n} \sum_{\mathbf{g} \in \mathcal{G}_p} f(\mathbf{g}, \mathcal{G}_\pi); \quad \text{Recall}(\mathcal{G}_\pi, \mathcal{G}_p) = \frac{1}{n} \sum_{\mathbf{g} \in \mathcal{G}_\pi} f(\mathbf{g}, \mathcal{G}_p).$$

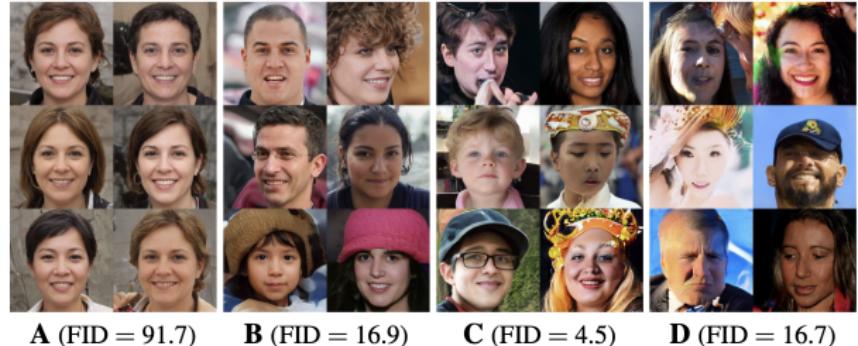
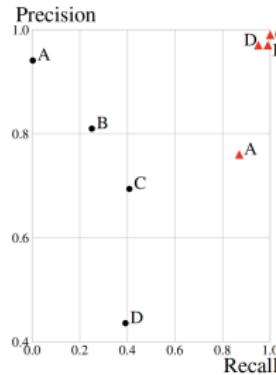


(a) True manifold



(b) Approx. manifold

Precision-Recall for generative models

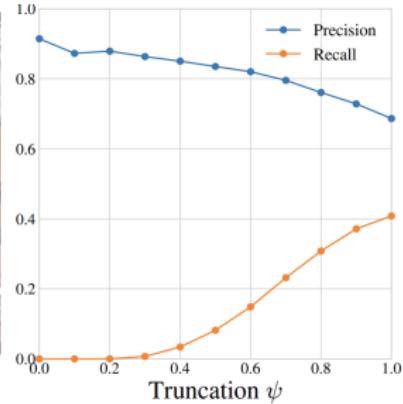


A (FID = 91.7)

B (FID = 16.9)

C (FID = 4.5)

D (FID = 16.7)



Truncation trick

BigGAN: truncated normal sampling

$$p(\mathbf{z}|\psi) = \mathcal{N}(\mathbf{z}|0, \mathbf{I}) / \int_{-\infty}^{\psi} \mathcal{N}(\mathbf{z}|0, \mathbf{I}) d\mathbf{z}$$

Components of $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ which fall outside a predefined range are resampled.

StyleGAN

$$\mathbf{z}' = \hat{\mathbf{z}} + \psi \cdot (\mathbf{z} - \hat{\mathbf{z}}), \quad \hat{\mathbf{z}} = \mathbb{E}_{\mathbf{z}} \mathbf{z}$$

- ▶ Constant ψ is a tradeoff between diversity and fidelity.
- ▶ $\psi = 0.7$ is used for most of the results.

Brock A., Donahue J., Simonyan K. Large Scale GAN Training for High Fidelity Natural Image Synthesis, 2018

Karras T., Laine S., Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks, 2018

Summary

- ▶ 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.
- ▶ Inception Score and Frechet Inception Distance are the common metrics for GAN evaluation, but both of them have drawbacks.
- ▶ Precision-recall allow to select model that compromises the sample quality and the sample diversity.
- ▶ Truncation tricks help to select model with compromised samples: diverse and sharp.