Deep Generative Models

Lecture 7

Roman Isachenko

AI Masters

2025, Spring

Recap of previous lecture

Likelihood-free learning

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

Imagine we have two sets of samples

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

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

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 $\mathbf{x}$.

- **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.

Recap of previous lecture

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.

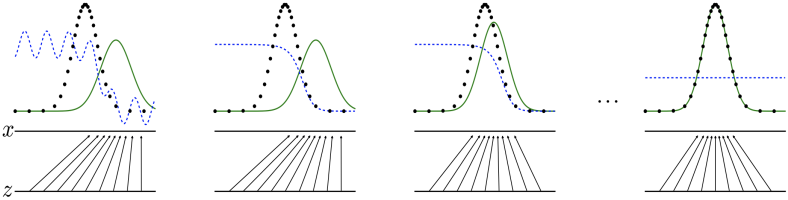
Goodfellow I. J. et al. Generative Adversarial Networks, 2014

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

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

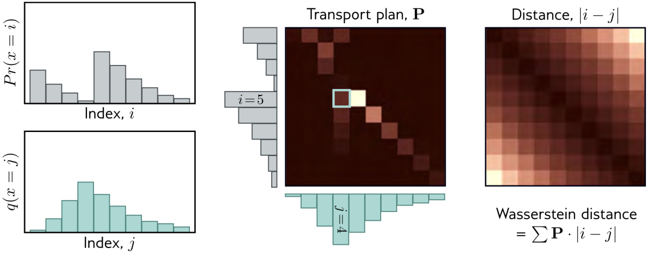
Informal theoretical results

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.

Simon J.D. Prince. Understanding Deep Learning, 2023

Outline

1. f-divergence minimization

2. Evaluation of likelihood-free models
   - Frechet Inception Distance (FID)
   - Precision-Recall

Outline

1. f-divergence minimization

2. Evaluation of likelihood-free models
   - Frechet Inception Distance (FID)
   - Precision-Recall

Divergences

- Forward KL divergence in the maximum likelihood estimation.
- Reverse KL in the variational inference (KL term in ELBO).
- JS divergence in the 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 $f(1) = 0$.

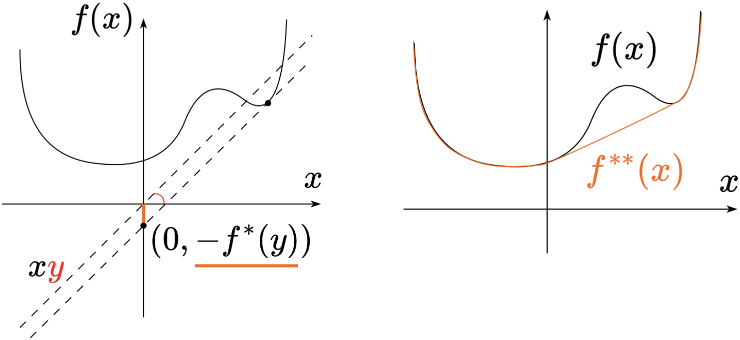
Name	$D_f(P\|Q)$	Generator $f(u)$
Kullback-Leibler	$\int p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x$	$u \log u$
Reverse KL	$\int q(x) \log \frac{q(x)}{p(x)} \, \mathrm{d}x$	$-\log u$
Pearson χ^2	$\int \frac{(q(x)-p(x))^2}{p(x)} \, \mathrm{d}x$	$(u-1)^2$
Squared Hellinger	$\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2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

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



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

---

image credit: https://fmin.xyz/docs/theory/Conjugate%20function.html

f-divergence family

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

Variational f-divergence estimation

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

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

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

$$\geq \sup_{T \in \mathcal{T}} \int \left(\pi(\mathbf{x}) T(\mathbf{x}) - p(\mathbf{x}) f^*(T(\mathbf{x})) \right) d\mathbf{x} =$$

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

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

f-divergence family

Variational divergence estimation

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

- Here $\mathcal{T}$ is a predefined class of functions.
- 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 $T(\mathbf{x}) = \log D(\mathbf{x})$ $(D(\mathbf{x}) \in [0, 1])$.

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

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 the lower bound we only need samples from $\pi(\mathbf{x})$ and $p(\mathbf{x})$. Hence, we are able to fit the 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. f-divergence minimization

2. Evaluation of likelihood-free models
   - Frechet Inception Distance (FID)
   - Precision-Recall

Evaluation of likelihood-free models

Likelihood-based models

- **train part:** fit the model.
- **validation part:** tune the hyperparameters.
- **test part:** evaluate generalization by reporting the likelihood.

Not all models have tractable likelihood
(VAE: compare ELBO values; GAN: **???**).

What do we want from samples?

- Sharpness

Low sharpness

High sharpness

- Diversity

Low diversity

High diversity

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

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.** The **conditional** distribution $p(y|\mathbf{x})$ should have low entropy (each image $\mathbf{x}$ depicts distinctly recognizable object).
- **Diversity.** The **marginal** distribution $p(y) = \int p(y|\mathbf{x})p(\mathbf{x})d\mathbf{x}$ should have high entropy (we generate all classes uniformly).

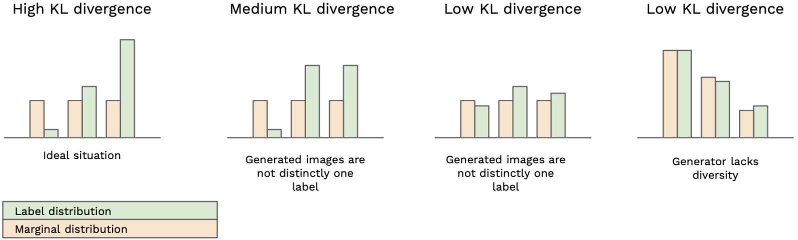


image credit: https://medium.com/octavian-ai/a-simple-explanation-of-the-inception-score-372dff6a8c7a

Outline

1. f-divergence minimization

2. Evaluation of likelihood-free models
   - Frechet Inception Distance (FID)
   - Precision-Recall

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

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

Frechet Inception Distance

$$\mathrm{FID}(\pi, p) = W_2^2(\pi, p)$$

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

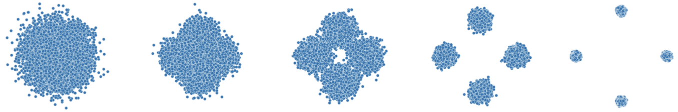
Heusel M. et al. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, 2017

Frechet Inception Distance (FID)

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

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

$FID(p(\mathbf{x}), \mathcal{N}(0, \mathbf{I}))$



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

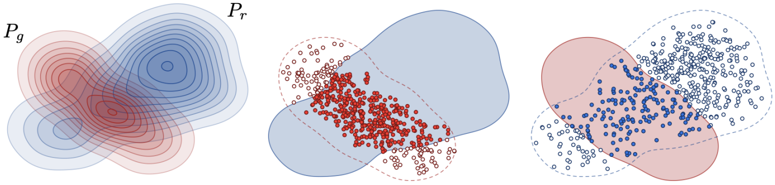
Outline

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

Precision-Recall

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.

Kynkäänniemi T. et al. Improved precision and recall metric for assessing generative models, 2019

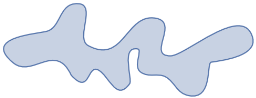
Precision-Recall

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

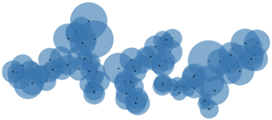
Define binary function:

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

$$\text{Precision}(\mathcal{S}_\pi, \mathcal{S}_p) = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{S}_p} \mathbb{I}(\mathbf{x}, \mathcal{S}_\pi); \;\; \text{Recall}(\mathcal{S}_\pi, \mathcal{S}_p) = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{S}_\pi} \mathbb{I}(\mathbf{x}, \mathcal{S}_p).$$



(a) True manifold

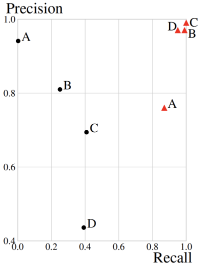


(b) Approx. manifold

Embed the samples using the pretrained network (as for FID).

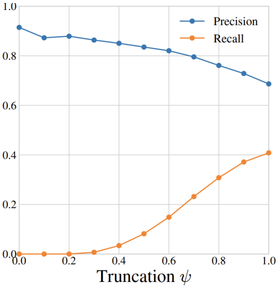
Kynkäänniemi T. et al. Improved precision and recall metric for assessing generative models, 2019

Precision-Recall



A (FID = 91.7) **B** (FID = 16.9) **C** (FID = 4.5) **D** (FID = 16.7)

$\psi = 0.0$ $\psi = 0.3$ $\psi = 0.7$ $\psi = 1.0$



Kynkäänniemi T. et al. Improved precision and recall metric for assessing generative models, 2019

Truncation trick

BigGAN: truncated normal sampling

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

Elements 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.

- Frechet Inception Distance is the most popular metric for the implicit models evaluation.

- Precision-recall allow to select model that compromises the sample quality and the sample diversity.

- Truncation tricks help to select model with the compromised samples: diverse and sharp.