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.

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

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 $\mathbf{x}$.

- **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

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

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.

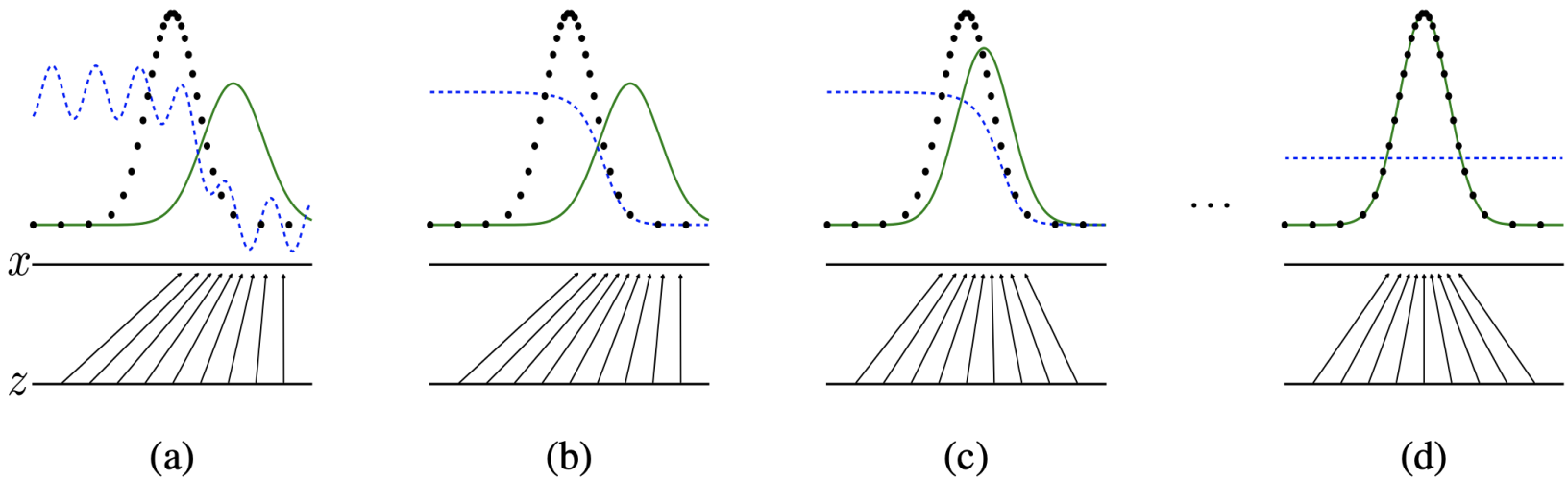
Goodfellow I. J. et al. Generative Adversarial Networks, 2014

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

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



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

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

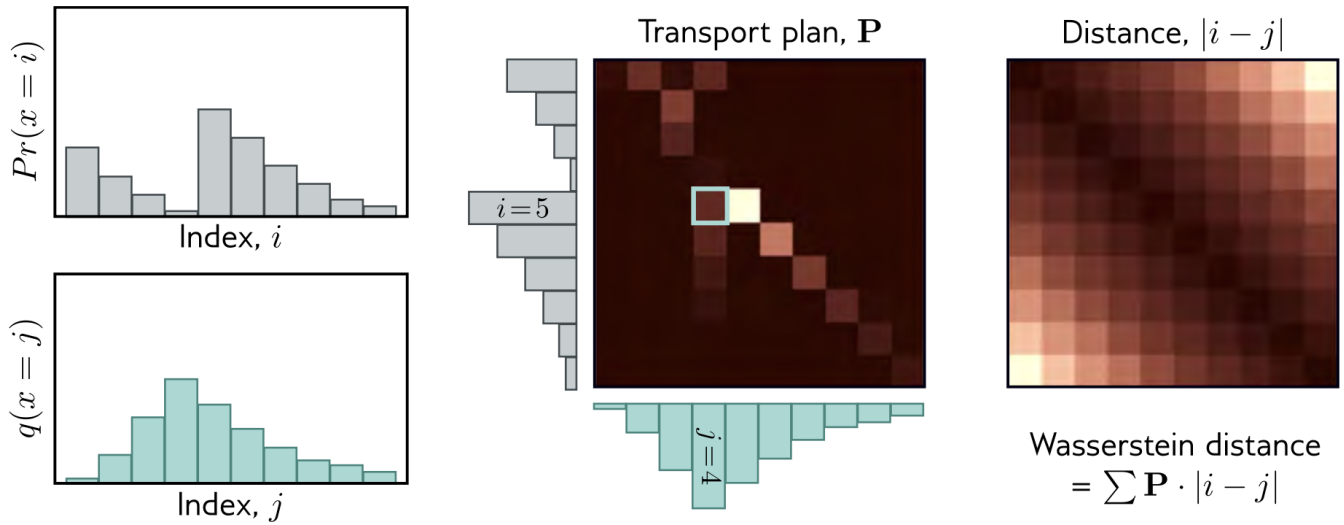
Informal theoretical results

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

Recap of previous lecture

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.

WGAN objective

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

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

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

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

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

Outline

1. Evaluation of likelihood-free models
   - Frechet Inception Distance (FID)
   - Precision-Recall
   - CLIP score
   - Human evaluation
2. Langevin dynamic
3. Score matching
4. Denoising score matching

Outline

1. Evaluation of likelihood-free models
   - Frechet Inception Distance (FID)
   - Precision-Recall
   - CLIP score
   - Human evaluation
2. Langevin dynamic
3. Score matching
4. Denoising score matching

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: **???**).

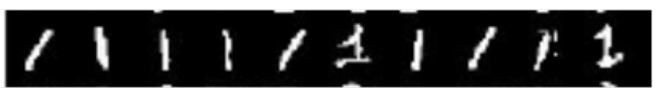
What do we want from samples?

- Sharpness

Low sharpness

High sharpness

- Diversity



Low diversity

High diversity

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

Outline

Wasserstein metric

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

Wassestein GAN (optimal transport)

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

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

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

Frechet Inception Distance

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

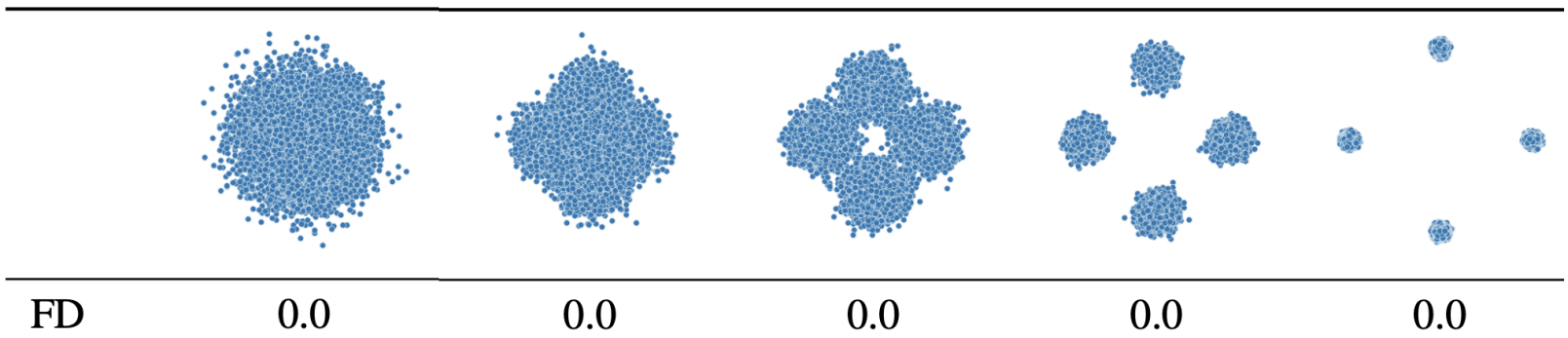
Heusel M. et al. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, 2017

Frechet Inception Distance (FID)

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

- FID is calculated in the latent space $\mathbf{z}$.
- We take pretrained image embedder to get the latent representations $\mathbf{z} = \mathbf{f}(\mathbf{x})$.
- $\boldsymbol{\mu}_\pi$, $\boldsymbol{\Sigma}_\pi$ and $\boldsymbol{\mu}_p$, $\boldsymbol{\Sigma}_p$ are the statistics of the latent representations $\mathbf{z}$ for the samples from $\pi(\mathbf{x})$ and $p(\mathbf{x}|\boldsymbol{\theta})$.

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



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

Frechet Inception Distance (FID)

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

Drawbacks

- Dependence on the pretrained classification model.
- Usage of the normality assumption.
- May not correspond to human evaluation.

Model	Model-A	Model-B
FID	21.40	18.42
FID_∞	20.16	17.19
Human rater preference	92.5%	6.9%

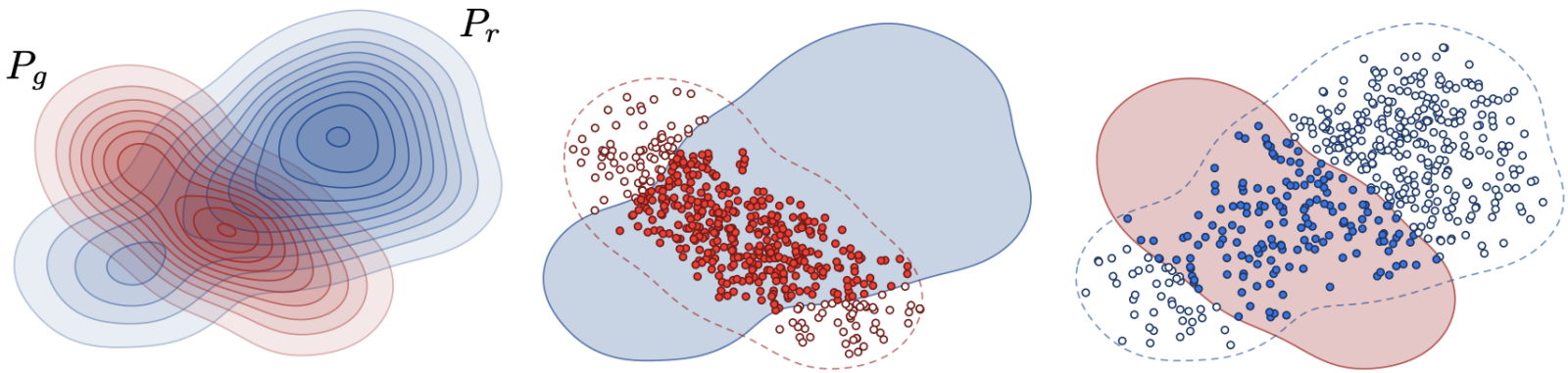
Jayasumana S. et al. Rethinking FID: Towards a Better Evaluation Metric for Image Generation, 2024

Outline

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

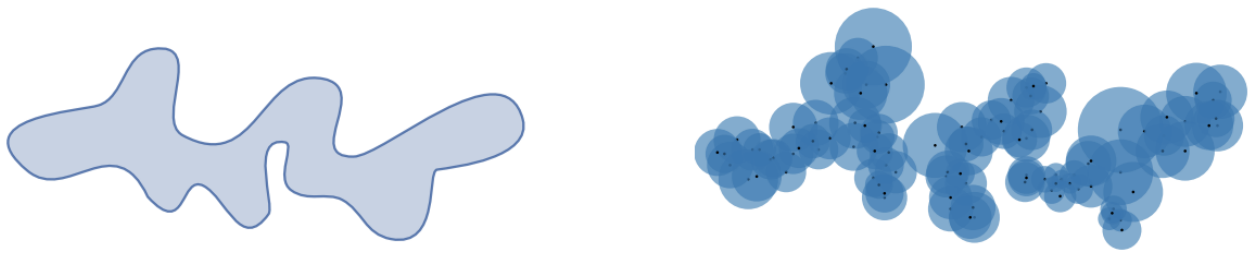
Precision-Recall

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

Define binary function:

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

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

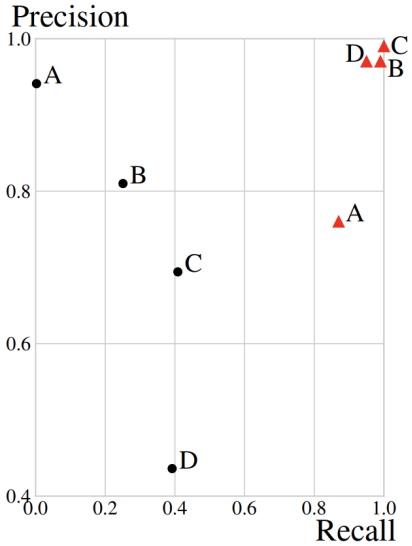


(a) True manifold (b) Approx. manifold

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

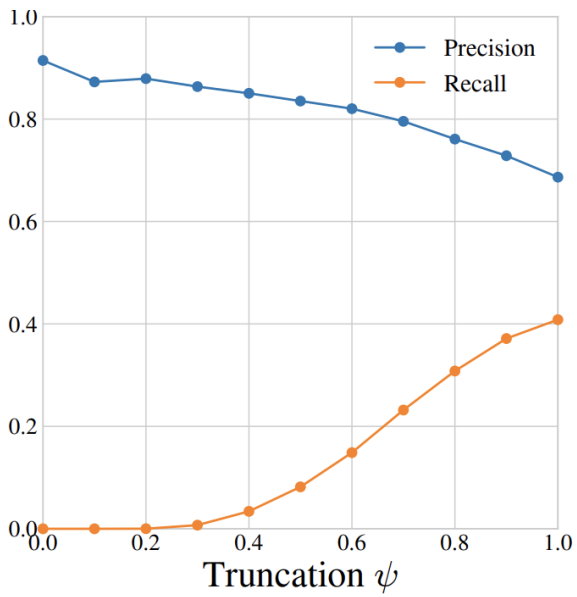
Kynkäänniemi T. et al. Improved precision and recall metric for assessing generative models, 2019

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.

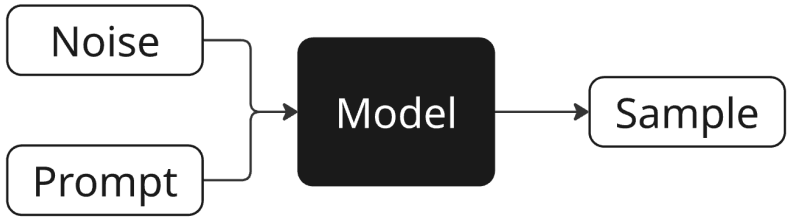
Brock A., Donahue J., Simonyan K. Large Scale GAN Training for High Fidelity Natural Image Synthesis, 2018
Karras T., Laine S., Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks, 2018

Outline

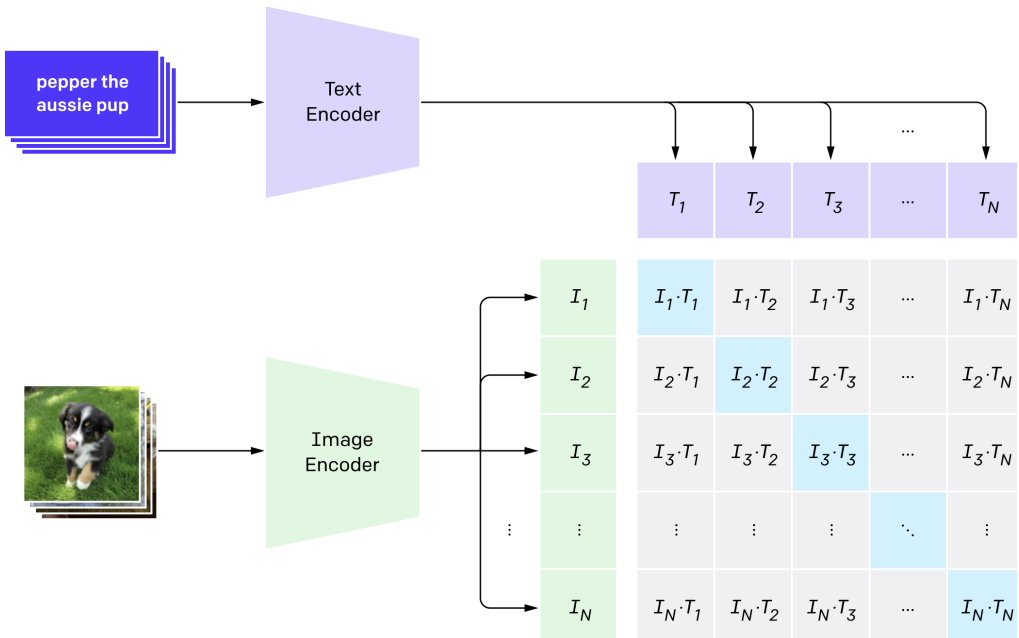
CLIP score

Unconditional model

Conditional model



We need the way to measure not only generated image quality, but also its relevance to the prompt.



Radford A. et al. Learning transferable visual models from natural language supervision, 2021

Outline

Human Evaluation

- There is no perfect automated metric.
- The best way to evaluate the generative model is to make human evaluation.
- It is essential to evaluate different aspects.

Аспект	Yandex ART 2.0	Mj 6.1	Mj 6	Ideogram	Recraft	Google Imagen3	Dall-E 3	FLUX	SBER Kandi3.1
Релевантность	0,59	0,58	0,63	**0,45**	**0,51**	**0,50**	**0,50**	**0,54**	0,75
Эстетика	**0,49**	0,55	0,55	**0,51**	**0,51**	0,61	0,61	**0,54**	0,59
Комплексность	**0,44**	0,73	0,70	0,68	0,76	0,75	0,75	0,71	0,74
Дефектность	0,69	0,57	0,68	0,55	0,59	0,63	0,63	**0,50**	0,75
Предпочтение	0,66	0,60	0,69	**0,49**	**0,54**	0,63	0,63	**0,51**	0,84

YandexART 2.5, 2025

Outline

Energy-based models

Unnormalized density

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{\hat{p}(\mathbf{x}|\boldsymbol{\theta})}{Z_{\boldsymbol{\theta}}}, \quad \text{where } Z_{\boldsymbol{\theta}} = \int \hat{p}(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x}$$

- $\hat{p}(\mathbf{x}|\boldsymbol{\theta})$ is any non-negative function.
- If we use the reparametrization $\hat{p}(\mathbf{x}|\boldsymbol{\theta}) = \exp(-f_{\boldsymbol{\theta}}(\mathbf{x}))$, we remove the non-negativite constraint.

Unnormalized density

The gradient of the normalized log-density equals to the gradient of the unnormalized log-density:

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}|\boldsymbol{\theta}) = \nabla_{\mathbf{x}} \log \hat{p}(\mathbf{x}|\boldsymbol{\theta}) - \nabla_{\mathbf{x}} \log Z_{\boldsymbol{\theta}} = \nabla_{\mathbf{x}} \log \hat{p}(\mathbf{x}|\boldsymbol{\theta})$$

- Let assume that we already have the density (normalized or unnormalized) $p(\mathbf{x}|\boldsymbol{\theta})$.
- How to sample from the model?

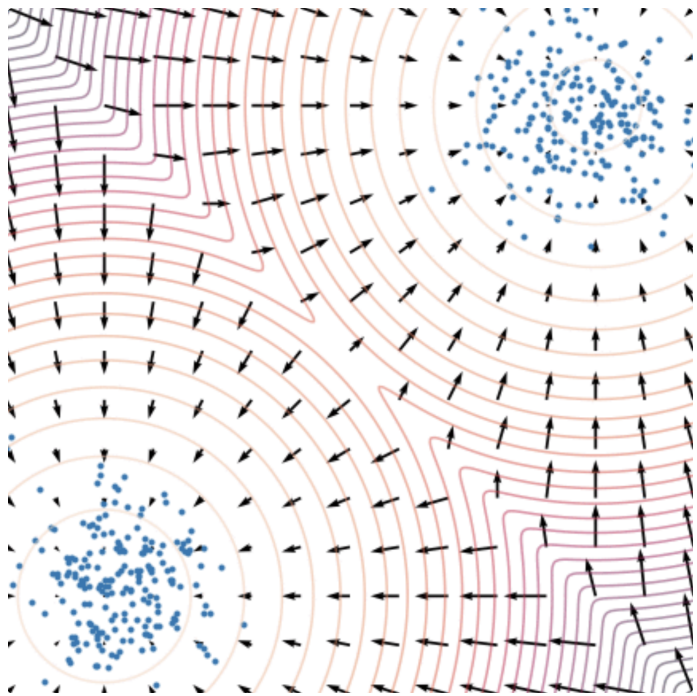
Langevin dynamic

Theorem (informal)

Let $\mathbf{x}_0$ be a random vector. Under some mild regularity conditions samples from the following dynamics will come from $p(\mathbf{x}|\boldsymbol{\theta})$ (for small enough η and large enough I)

$$\mathbf{x}_{I+1} = \mathbf{x}_I + \frac{\eta}{2} \cdot \nabla_{\mathbf{x}_I} \log p(\mathbf{x}_I|\boldsymbol{\theta}) + \sqrt{\eta} \cdot \boldsymbol{\epsilon}_I, \quad \boldsymbol{\epsilon}_I \sim \mathcal{N}(0, \mathbf{I}).$$

- What do we get if $\boldsymbol{\epsilon}_I = \mathbf{0}$?
- The density $p(\mathbf{x}|\boldsymbol{\theta})$ is a **stationary** distribution for the Markov chain.
- We take the gradient w.r.t. to $\mathbf{x}$ (not $\boldsymbol{\theta}$).
- $\nabla_{\mathbf{x}} \log p(\mathbf{x}|\boldsymbol{\theta})$ defines the vector field.



Song Y. Generative Modeling by Estimating Gradients of the Data Distribution, blog post, 2021

Outline

Score matching

Score function

$$\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}) = \nabla_{\mathbf{x}} \log p(\mathbf{x}|\boldsymbol{\theta})$$

Langevin dynamic

If we find the score function $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}) = \nabla_{\mathbf{x}} \log p(\mathbf{x}|\boldsymbol{\theta})$ we will be able to sample from the model using Langevin dynamic.

$$\mathbf{x}_{l+1} = \mathbf{x}_l + \frac{\eta}{2} \cdot \nabla_{\mathbf{x}_l} \log p(\mathbf{x}_l|\boldsymbol{\theta}) + \sqrt{\eta} \cdot \boldsymbol{\epsilon}_l = \mathbf{x}_l + \frac{\eta}{2} \cdot \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_l) + \sqrt{\eta} \cdot \boldsymbol{\epsilon}_l.$$

Fisher divergence

$$D_F(\pi, p) = \frac{1}{2}\mathbb{E}_\pi \left\| \nabla_{\mathbf{x}} \log p(\mathbf{x}|\boldsymbol{\theta}) - \nabla_{\mathbf{x}} \log \pi(\mathbf{x}) \right\|_2^2 =$$

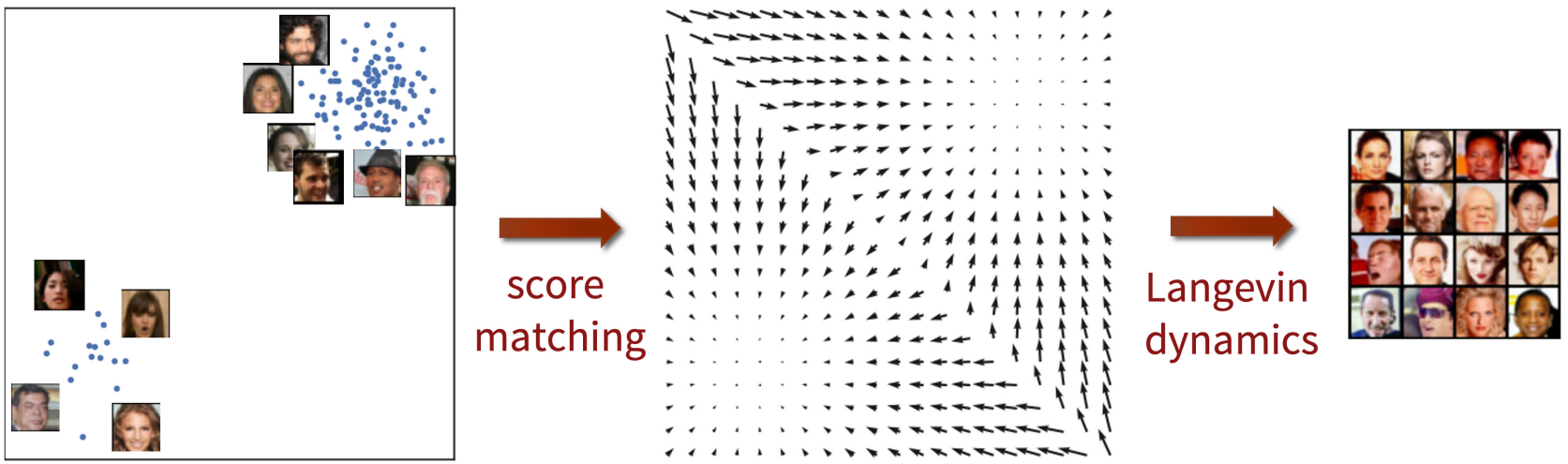
$$= \frac{1}{2}\mathbb{E}_\pi \left\| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}) - \nabla_{\mathbf{x}} \log \pi(\mathbf{x}) \right\|_2^2 \to \min_{\boldsymbol{\theta}}$$

Song Y. Generative Modeling by Estimating Gradients of the Data Distribution, blog post, 2021

Score matching

Fisher divergence

$$D_F(\pi, p) = \frac{1}{2}\mathbb{E}_\pi \left\| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}) - \nabla_{\mathbf{x}} \log \pi(\mathbf{x}) \right\|_2^2 \to \min_{\boldsymbol{\theta}}$$



Data samples

$\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \overset{\text{i.i.d.}}{\sim} p(\mathbf{x})$

Scores

$\mathbf{s}_\theta(\mathbf{x}) \approx \nabla_{\mathbf{x}} \log p(\mathbf{x})$

New samples

Problem: We do not know $\nabla_{\mathbf{x}} \log \pi(\mathbf{x})$.

Song Y. Generative Modeling by Estimating Gradients of the Data Distribution, blog post, 2021

Outline

Denoising score matching

Let perturb original data $\mathbf{x} \sim \pi(\mathbf{x})$ by random normal noise

$$\mathbf{x}_\sigma = \mathbf{x} + \sigma \cdot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \quad q(\mathbf{x}_\sigma|\mathbf{x}) = \mathcal{N}(\mathbf{x}, \sigma^2 \cdot \mathbf{I})$$

$$q(\mathbf{x}_\sigma) = \int q(\mathbf{x}_\sigma|\mathbf{x})\pi(\mathbf{x})d\mathbf{x}.$$

Assumption

The solution of

$$\frac{1}{2}\mathbb{E}_{q(\mathbf{x}_\sigma)}\left\|\mathbf{s}_{\boldsymbol{\theta},\sigma}(\mathbf{x}_\sigma) - \nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma)\right\|_2^2 \to \min_{\boldsymbol{\theta}}$$

satisfies $\mathbf{s}_{\boldsymbol{\theta},\sigma}(\mathbf{x}_\sigma) \approx \mathbf{s}_{\boldsymbol{\theta},0}(\mathbf{x}_0) = \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x})$ if σ is small enough.

- The score function of the noised data is almost the same as the score function of the original data.
- Score function $\mathbf{s}_{\boldsymbol{\theta},\sigma}(\mathbf{x}_\sigma)$ parametrized by σ.
- **Note:** We don't know $q(\mathbf{x}_\sigma)$, just like $\pi(\mathbf{x})$.

Vincent P. A Connection Between Score Matching and Denoising Autoencoders, 2010

Denoising score matching

Theorem

$$\mathbb{E}_{q(\mathbf{x}_\sigma)} \left\| \mathbf{s}_{\boldsymbol{\theta},\sigma}(\mathbf{x}_\sigma) - \nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma) \right\|_2^2 =$$
$$= \mathbb{E}_{\pi(\mathbf{x})} \mathbb{E}_{q(\mathbf{x}_\sigma|\mathbf{x})} \left\| \mathbf{s}_{\boldsymbol{\theta},\sigma}(\mathbf{x}_\sigma) - \nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma|\mathbf{x}) \right\|_2^2 + \text{const}(\boldsymbol{\theta})$$

Gradient of the noise kernel

$$\mathbf{x}_\sigma = \mathbf{x} + \sigma \cdot \boldsymbol{\epsilon}, \quad q(\mathbf{x}_\sigma|\mathbf{x}) = \mathcal{N}(\mathbf{x}, \sigma^2 \cdot \mathbf{I})$$

$$\nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma|\mathbf{x}) = -\frac{\mathbf{x}_\sigma - \mathbf{x}}{\sigma^2} = -\frac{\boldsymbol{\epsilon}}{\sigma}$$

- The RHS does not need to compute $\nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma)$ and even $\nabla_{\mathbf{x}_\sigma} \log \pi(\mathbf{x}_\sigma)$.
- $\mathbf{s}_{\boldsymbol{\theta},\sigma}(\mathbf{x}_\sigma)$ tries to **denoise** the noised samples $\mathbf{x}_\sigma$.

Vincent P. A Connection Between Score Matching and Denoising Autoencoders, 2010

Denoising score matching

Initial objective:

$$\mathbb{E}_{\pi(\mathbf{x})} \left\| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}) - \nabla_{\mathbf{x}} \log \pi(\mathbf{x}) \right\|_2^2 \to \min_{\boldsymbol{\theta}}$$

Noised objective:

$$\mathbb{E}_{q(\mathbf{x}_\sigma)} \left\| \mathbf{s}_{\boldsymbol{\theta},\sigma}(\mathbf{x}_\sigma) - \nabla_{\mathbf{x}} \log q(\mathbf{x}_\sigma) \right\|_2^2 \to \min_{\boldsymbol{\theta}}$$

This is equivalent to denoising task

$$\mathbb{E}_{\pi(\mathbf{x})} \mathbb{E}_{q(\mathbf{x}_\sigma|\mathbf{x})} \left\| \mathbf{s}_{\boldsymbol{\theta},\sigma}(\mathbf{x}_\sigma) - \nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma|\mathbf{x}) \right\|_2^2 \to \min_{\boldsymbol{\theta}}$$

$$\mathbb{E}_{\pi(\mathbf{x})} \mathbb{E}_{\mathcal{N}(0,\mathbf{I})} \left\| \mathbf{s}_{\boldsymbol{\theta},\sigma}(\mathbf{x} + \sigma \cdot \boldsymbol{\epsilon}) + \frac{\boldsymbol{\epsilon}}{\sigma} \right\|_2^2 \to \min_{\boldsymbol{\theta}}$$

Langevin dynamic

$$\mathbf{x}_{l+1} = \mathbf{x}_l + \frac{\eta}{2} \cdot \mathbf{s}_{\boldsymbol{\theta},\sigma}(\mathbf{x}_l) + \sqrt{\eta} \cdot \boldsymbol{\epsilon}_l, \quad \boldsymbol{\epsilon}_l \sim \mathcal{N}(0, \mathbf{I}).$$

Song Y. Generative Modeling by Estimating Gradients of the Data Distribution, blog post, 2021

Summary

- 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.
- CLIP score is frequently used to measure text-to-image relevance.
- The best way to measure the generated image quality is to make human evaluation.
- Langevin dynamics allows to sample from the generative model using the gradient of the log-likelihood.
- Score matching proposes to minimize the Fisher divergence to get the score function.
- Denoising score matching minimizes the Fisher divergence on noisy samples. It allows to estimate the Fisher divergence using samples.