

# Deep Generative Models

## Lecture 6

Roman Isachenko

Moscow Institute of Physics and Technology  
Yandex School of Data Analysis

2025, Autumn

# Recap of Previous Lecture

## Likelihood-Free Learning

- ▶ Likelihood isn't a perfect metric for generative models.
- ▶ Likelihood may be intractable.

Imagine we have two sets of samples:

- ▶  $\{\mathbf{x}_i\}_{i=1}^{n_1} \sim p_{\text{data}}(\mathbf{x})$  – real samples;
- ▶  $\{\mathbf{x}_i\}_{i=1}^{n_2} \sim p_{\theta}(\mathbf{x})$  – generated (fake) samples.

$$p(y = 1|\mathbf{x}) = P(\mathbf{x} \sim p_{\text{data}}(\mathbf{x})); \quad p(y = 0|\mathbf{x}) = P(\mathbf{x} \sim p_{\theta}(\mathbf{x}))$$

## Assumption

The generative distribution  $p_{\theta}(\mathbf{x})$  matches the true distribution  $p_{\text{data}}(\mathbf{x})$  if we can't distinguish between them using a discriminative model  $p(y|\mathbf{x})$ .

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

## Recap of Previous Lecture

### GAN Optimality Theorem

The minimax game

$$\min_G \max_D \underbrace{\left[ \mathbb{E}_{p_{\text{data}}(\mathbf{x})} \log D(\mathbf{x}) + \mathbb{E}_{p(\mathbf{z})} \log(1 - D(\mathbf{G}(\mathbf{z}))) \right]}_{V(G,D)}$$

has a global optimum at  $p_{\text{data}}(\mathbf{x}) = p_\theta(\mathbf{x})$ , and then  $D^*(\mathbf{x}) = 0.5$ .

$$\min_G V(G, D^*) = \min_G [2\text{JSD}(\pi \| p) - \log 4] = -\log 4, \quad p_{\text{data}}(\mathbf{x}) = p_\theta(\mathbf{x}).$$

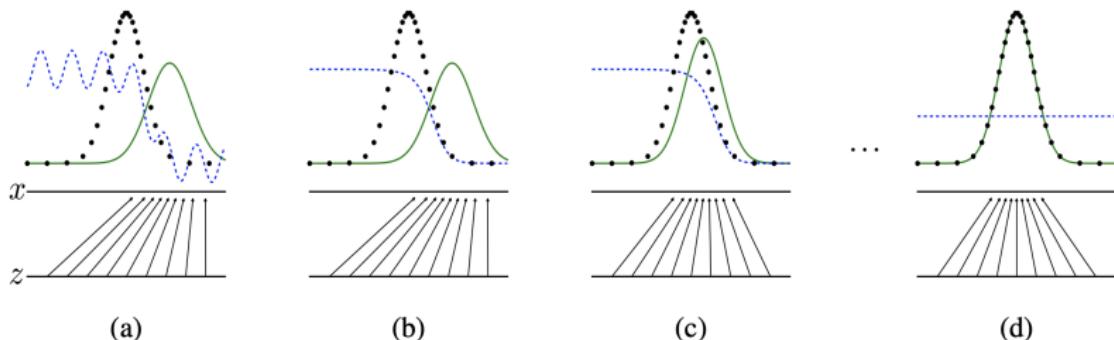
If the generator can be **any** function and the discriminator is **optimal** at each step, then the generator is **guaranteed to converge** to the data distribution.

# Recap of Previous Lecture

- ▶ The generator is updated in the parameter space; the discriminator isn't optimal at every iteration.
- ▶ Both generator and discriminator loss typically oscillate during GAN training.

## Objective

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



# Recap of Previous Lecture

## Main Issues With Standard GANs

- ▶ Vanishing gradients (solution: non-saturating GAN).
- ▶ Mode collapse (arises from Jensen-Shannon divergence).

## Standard GAN

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

## Informal Theoretical Results

Both the data distribution  $p_{\text{data}}(\mathbf{x})$  and the generative distribution  $p_{\theta}(\mathbf{x})$  are low-dimensional with disjoint supports. In such cases,

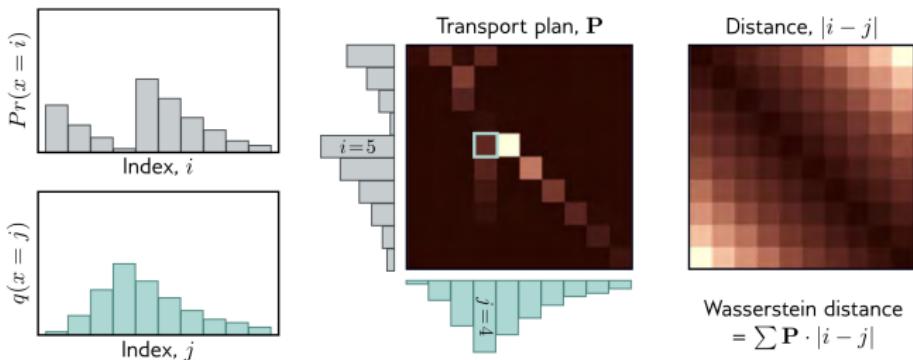
$$\text{KL}(p_{\text{data}} \| p_{\theta}) = \text{KL}(p_{\theta} \| p_{\text{data}}) = \infty, \quad \text{JSD}(p_{\text{data}} \| p_{\theta}) = \log 2.$$

---

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

$$W(\pi \| p) = \inf_{\gamma \in \Gamma(\pi, p)} \mathbb{E}_{(\mathbf{x}_1, \mathbf{x}_2) \sim \gamma} \|\mathbf{x}_1 - \mathbf{x}_2\| = \inf_{\gamma \in \Gamma(\pi, p)} \int \|\mathbf{x}_1 - \mathbf{x}_2\| \gamma(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_1 d\mathbf{x}_2$$

- ▶  $\gamma(\mathbf{x}_1, \mathbf{x}_2)$  – transportation plan (amount of "dirt" to transport from  $\mathbf{x}_1$  to  $\mathbf{x}_2$ ).
- ▶  $\Gamma(\pi, p)$  – set of all joint distributions  $\gamma(\mathbf{x}_1, \mathbf{x}_2)$  with marginals  $\pi$  and  $p$  ( $\int \gamma(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_1 = p(\mathbf{x}_2)$ ,  $\int \gamma(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_2 = \pi(\mathbf{x}_1)$ ).
- ▶  $\gamma(\mathbf{x}_1, \mathbf{x}_2)$  – the amount;  $\|\mathbf{x}_1 - \mathbf{x}_2\|$  – the distance.

# Recap of Previous Lecture

## Theorem (Kantorovich-Rubinstein Duality)

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

where  $\|f\|_L \leq K$  denotes  $K$ -Lipschitz continuous functions.

## WGAN Objective

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

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

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

# Outline

## 1. Evaluation of Likelihood-Free Models

Frechet Inception Distance (FID)

Precision-Recall

CLIP Score

Human Evaluation

## 2. Langevin Dynamics

## 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 Dynamics

## 3. Score Matching

## 4. Denoising Score Matching

# Evaluation of Likelihood-Free Models

## Likelihood-Based Models

- ▶ **Train:** fit the model.
- ▶ **Validation:** tune hyperparameters.
- ▶ **Test:** assess generalization by reporting likelihood.

# Evaluation of Likelihood-Free Models

## Likelihood-Based Models

- ▶ **Train:** fit the model.
- ▶ **Validation:** tune hyperparameters.
- ▶ **Test:** assess generalization by reporting likelihood.

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

# Evaluation of Likelihood-Free Models

## Likelihood-Based Models

- ▶ **Train:** fit the model.
- ▶ **Validation:** tune hyperparameters.
- ▶ **Test:** assess generalization by reporting likelihood.

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

## Desirable Properties for Samples

- ▶ Sharpness



# Evaluation of Likelihood-Free Models

## Likelihood-Based Models

- ▶ **Train:** fit the model.
- ▶ **Validation:** tune hyperparameters.
- ▶ **Test:** assess generalization by reporting likelihood.

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

## Desirable Properties for Samples

- ▶ Sharpness



- ▶ Diversity



# Outline

## 1. Evaluation of Likelihood-Free Models

Frechet Inception Distance (FID)

Precision-Recall

CLIP Score

Human Evaluation

## 2. Langevin Dynamics

## 3. Score Matching

## 4. Denoising Score Matching

## Wasserstein Metric

$$W_s(\pi \| p) = \inf_{\gamma \in \Gamma(\pi, p)} \left( \mathbb{E}_{(\mathbf{x}_1, \mathbf{x}_2) \sim \gamma} \|\mathbf{x}_1 - \mathbf{x}_2\|^s \right)^{1/s}$$

## Wasserstein Metric

$$W_s(\pi \| p) = \inf_{\gamma \in \Gamma(\pi, p)} \left( \mathbb{E}_{(\mathbf{x}_1, \mathbf{x}_2) \sim \gamma} \|\mathbf{x}_1 - \mathbf{x}_2\|^s \right)^{1/s}$$

## Wasserstein GAN (Optimal Transport)

$$W(\pi \| p) = \inf_{\gamma \in \Gamma(\pi, p)} \mathbb{E}_{(\mathbf{x}_1, \mathbf{x}_2) \sim \gamma} \|\mathbf{x}_1 - \mathbf{x}_2\| = \inf_{\gamma \in \Gamma(\pi, p)} \int \|\mathbf{x}_1 - \mathbf{x}_2\| \gamma(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_1 d\mathbf{x}_2$$

# Wasserstein Metric

$$W_s(\pi \| p) = \inf_{\gamma \in \Gamma(\pi, p)} (\mathbb{E}_{(\mathbf{x}_1, \mathbf{x}_2) \sim \gamma} \|\mathbf{x}_1 - \mathbf{x}_2\|^s)^{1/s}$$

## Wasserstein GAN (Optimal Transport)

$$W(\pi \| p) = \inf_{\gamma \in \Gamma(\pi, p)} \mathbb{E}_{(\mathbf{x}_1, \mathbf{x}_2) \sim \gamma} \|\mathbf{x}_1 - \mathbf{x}_2\| = \inf_{\gamma \in \Gamma(\pi, p)} \int \|\mathbf{x}_1 - \mathbf{x}_2\| \gamma(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_1 d\mathbf{x}_2$$

### Theorem

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

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

## Wasserstein Metric

$$W_s(\pi \| p) = \inf_{\gamma \in \Gamma(\pi, p)} (\mathbb{E}_{(\mathbf{x}_1, \mathbf{x}_2) \sim \gamma} \|\mathbf{x}_1 - \mathbf{x}_2\|^s)^{1/s}$$

## Wasserstein GAN (Optimal Transport)

$$W(\pi \| p) = \inf_{\gamma \in \Gamma(\pi, p)} \mathbb{E}_{(\mathbf{x}_1, \mathbf{x}_2) \sim \gamma} \|\mathbf{x}_1 - \mathbf{x}_2\| = \inf_{\gamma \in \Gamma(\pi, p)} \int \|\mathbf{x}_1 - \mathbf{x}_2\| \gamma(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_1 d\mathbf{x}_2$$

## Theorem

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

$$\text{FID}(p_{\text{data}}, p_\theta) = W_2^2(p_{\text{data}} \| p_\theta)$$

## Frechet Inception Distance (FID)

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

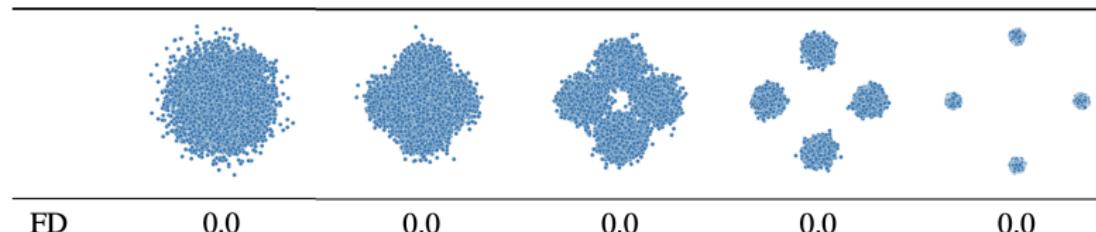
- ▶ FID is computed in the latent space  $\mathbf{z}$ .
- ▶ We use a pretrained image embedder to get latent representations  $\mathbf{z} = \mathbf{f}(\mathbf{x})$ .
- ▶  $\boldsymbol{\mu}_{\text{data}}$ ,  $\boldsymbol{\Sigma}_{\text{data}}$  and  $\boldsymbol{\mu}_{\theta}$ ,  $\boldsymbol{\Sigma}_{\theta}$  are statistics of latent representations for samples from  $p_{\text{data}}(\mathbf{x})$  and  $p_{\theta}(\mathbf{x})$ .

## Frechet Inception Distance (FID)

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

- ▶ FID is computed in the latent space  $\mathbf{z}$ .
- ▶ We use a pretrained image embedder to get latent representations  $\mathbf{z} = \mathbf{f}(\mathbf{x})$ .
- ▶  $\boldsymbol{\mu}_{\text{data}}$ ,  $\boldsymbol{\Sigma}_{\text{data}}$  and  $\boldsymbol{\mu}_{\theta}$ ,  $\boldsymbol{\Sigma}_{\theta}$  are statistics of latent representations for samples from  $p_{\text{data}}(\mathbf{x})$  and  $p_{\theta}(\mathbf{x})$ .

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



## Frechet Inception Distance (FID)

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

## Frechet Inception Distance (FID)

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

### Drawbacks

- ▶ Depends on the pretrained classification network.
- ▶ Uses the normality assumption.
- ▶ May not correlate with human evaluation.

Model	Model-A	Model-B
FID	21.40	18.42
$\text{FID}_{\infty}$	20.16	17.19
Human rater preference	92.5%	6.9%

# Outline

## 1. Evaluation of Likelihood-Free Models

Frechet Inception Distance (FID)

Precision-Recall

CLIP Score

Human Evaluation

## 2. Langevin Dynamics

## 3. Score Matching

## 4. Denoising Score Matching

# Precision-Recall

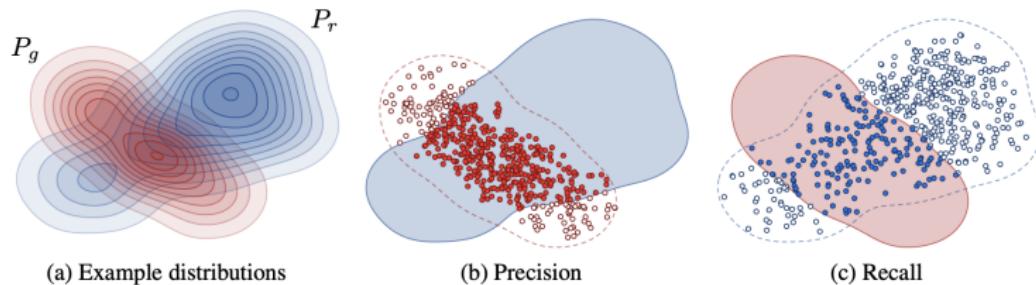
## Desirable Properties for Samples

- ▶ **Sharpness:** generated samples should possess high visual quality.
- ▶ **Diversity:** their variation should match that in the training data.

# Precision-Recall

## Desirable Properties for Samples

- ▶ **Sharpness:** generated samples should possess high visual quality.
- ▶ **Diversity:** their variation should match that in the training data.



- ▶ **Precision** denotes the fraction of generated images that look realistic.
- ▶ **Recall** measures how well the generator covers the training data manifold.

## Precision-Recall

- ▶  $\mathcal{S}_{\text{data}} = \{\mathbf{x}_i\}_{i=1}^n \sim p_{\text{data}}(\mathbf{x})$  – real samples;
- ▶  $\mathcal{S}_{\theta} = \{\mathbf{x}_i\}_{i=1}^n \sim p_{\theta}(\mathbf{x})$  – generated samples.

## Precision-Recall

- ▶  $\mathcal{S}_{\text{data}} = \{\mathbf{x}_i\}_{i=1}^n \sim p_{\text{data}}(\mathbf{x})$  – real samples;
- ▶  $\mathcal{S}_{\theta} = \{\mathbf{x}_i\}_{i=1}^n \sim p_{\theta}(\mathbf{x})$  – generated samples.

Define a binary function:

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

## Precision-Recall

- ▶  $\mathcal{S}_{\text{data}} = \{\mathbf{x}_i\}_{i=1}^n \sim p_{\text{data}}(\mathbf{x})$  – real samples;
- ▶  $\mathcal{S}_{\theta} = \{\mathbf{x}_i\}_{i=1}^n \sim p_{\theta}(\mathbf{x})$  – generated samples.

Define a binary function:

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

$$\Pr(\mathcal{S}_{\text{data}}, \mathcal{S}_{\theta}) = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{S}_{\theta}} \mathbb{I}(\mathbf{x}, \mathcal{S}_{\text{data}}); \quad \text{Rec}(\mathcal{S}_{\text{data}}, \mathcal{S}_{\theta}) = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{S}_{\text{data}}} \mathbb{I}(\mathbf{x}, \mathcal{S}_{\theta}).$$

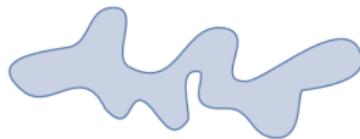
## Precision-Recall

- ▶  $\mathcal{S}_{\text{data}} = \{\mathbf{x}_i\}_{i=1}^n \sim p_{\text{data}}(\mathbf{x})$  – real samples;
- ▶  $\mathcal{S}_{\theta} = \{\mathbf{x}_i\}_{i=1}^n \sim p_{\theta}(\mathbf{x})$  – generated samples.

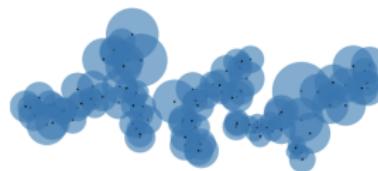
Define a binary function:

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

$$\Pr(\mathcal{S}_{\text{data}}, \mathcal{S}_{\theta}) = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{S}_{\theta}} \mathbb{I}(\mathbf{x}, \mathcal{S}_{\text{data}}); \quad \text{Rec}(\mathcal{S}_{\text{data}}, \mathcal{S}_{\theta}) = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{S}_{\text{data}}} \mathbb{I}(\mathbf{x}, \mathcal{S}_{\theta}).$$



(a) True manifold



(b) Approx. manifold

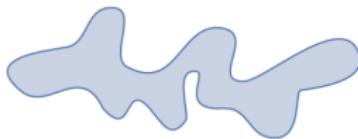
# Precision-Recall

- ▶  $\mathcal{S}_{\text{data}} = \{\mathbf{x}_i\}_{i=1}^n \sim p_{\text{data}}(\mathbf{x})$  – real samples;
- ▶  $\mathcal{S}_{\theta} = \{\mathbf{x}_i\}_{i=1}^n \sim p_{\theta}(\mathbf{x})$  – generated samples.

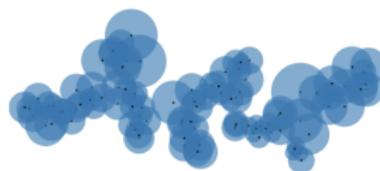
Define a binary function:

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

$$\Pr(\mathcal{S}_{\text{data}}, \mathcal{S}_{\theta}) = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{S}_{\theta}} \mathbb{I}(\mathbf{x}, \mathcal{S}_{\text{data}}); \quad \text{Rec}(\mathcal{S}_{\text{data}}, \mathcal{S}_{\theta}) = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{S}_{\text{data}}} \mathbb{I}(\mathbf{x}, \mathcal{S}_{\theta}).$$



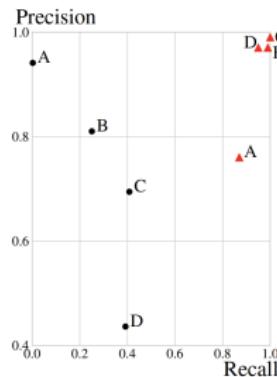
(a) True manifold



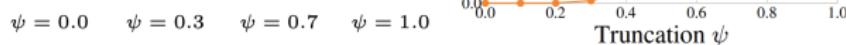
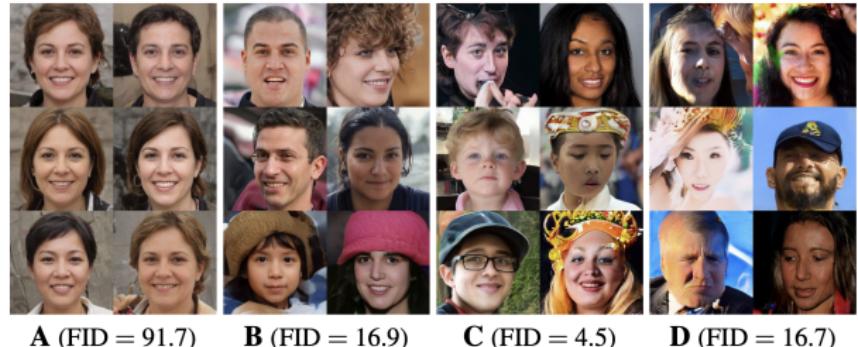
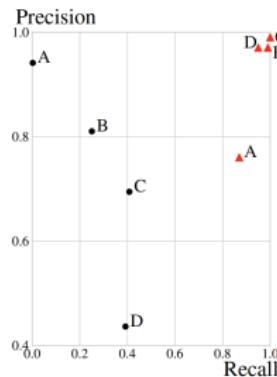
(b) Approx. manifold

Embed the samples using a pretrained network (as in FID).

# Precision-Recall



# Precision-Recall



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

# Outline

## 1. Evaluation of Likelihood-Free Models

Frechet Inception Distance (FID)

Precision-Recall

**CLIP Score**

Human Evaluation

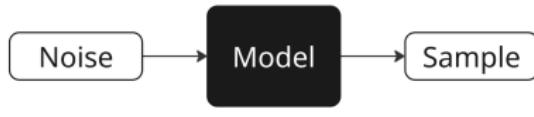
## 2. Langevin Dynamics

## 3. Score Matching

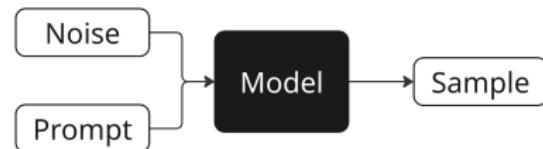
## 4. Denoising Score Matching

# CLIP Score

## Unconditional Model

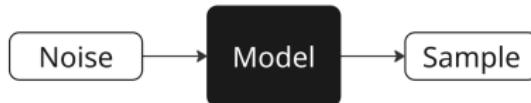


## Conditional Model

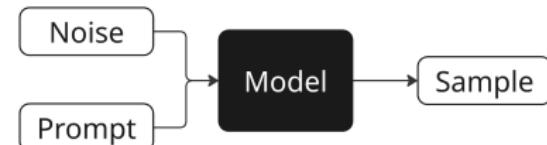


# CLIP Score

## Unconditional Model



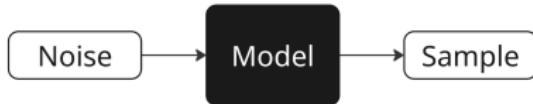
## Conditional Model



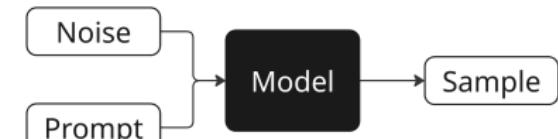
We need a way to measure not only the quality of the generated image, but also how well it's aligned with the prompt.

# CLIP Score

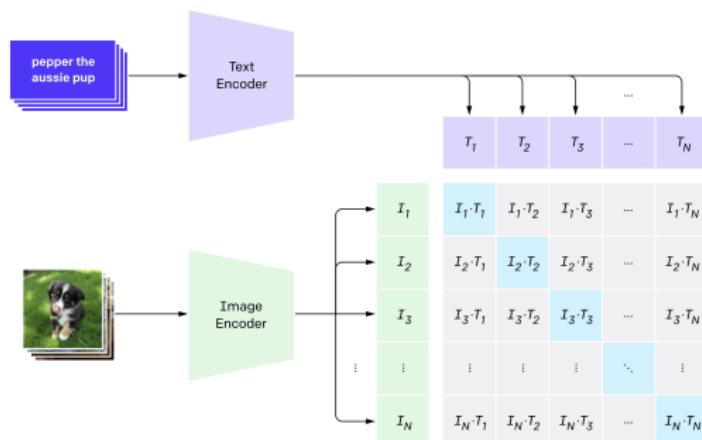
## Unconditional Model



## Conditional Model



We need a way to measure not only the quality of the generated image, but also how well it's aligned with the prompt.



# Outline

## 1. Evaluation of Likelihood-Free Models

Frechet Inception Distance (FID)

Precision-Recall

CLIP Score

Human Evaluation

## 2. Langevin Dynamics

## 3. Score Matching

## 4. Denoising Score Matching

## Human Evaluation

- ▶ No automated metric is perfect.
- ▶ The best way to evaluate generative models is by human assessment.
- ▶ It's important to assess various properties.

# Human Evaluation

- ▶ No automated metric is perfect.
- ▶ The best way to evaluate generative models is by human assessment.
- ▶ It's important to assess various properties.

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

# Outline

## 1. Evaluation of Likelihood-Free Models

Frechet Inception Distance (FID)

Precision-Recall

CLIP Score

Human Evaluation

## 2. Langevin Dynamics

## 3. Score Matching

## 4. Denoising Score Matching

# Energy-Based Models

## Unnormalized Density

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

- ▶  $\hat{p}_{\theta}(\mathbf{x})$  can be any non-negative function.
- ▶ If we reparameterize as  $\hat{p}_{\theta}(\mathbf{x}) = \exp(-f_{\theta}(\mathbf{x}))$ , we eliminate the non-negativity constraint.

# Energy-Based Models

## Unnormalized Density

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

- ▶  $\hat{p}_{\theta}(\mathbf{x})$  can be any non-negative function.
- ▶ If we reparameterize as  $\hat{p}_{\theta}(\mathbf{x}) = \exp(-f_{\theta}(\mathbf{x}))$ , we eliminate the non-negativity constraint.

## Unnormalized Density

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

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

# Energy-Based Models

## Unnormalized Density

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

- ▶  $\hat{p}_{\theta}(\mathbf{x})$  can be any non-negative function.
- ▶ If we reparameterize as  $\hat{p}_{\theta}(\mathbf{x}) = \exp(-f_{\theta}(\mathbf{x}))$ , we eliminate the non-negativity constraint.

## Unnormalized Density

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

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

- ▶ Suppose we already have this density (normalized or not)  $p_{\theta}(\mathbf{x})$ .
- ▶ How can we sample from the model?

# Langevin Dynamics

## Theorem

Consider energy-based model  $p(\mathbf{x}) = \frac{\hat{p}(\mathbf{x})}{Z}$ ,  $\hat{p}(\mathbf{x}) = \exp(-f(\mathbf{x}))$ , with continuously differentiable  $f(\mathbf{x}) : \mathbb{R}^m \rightarrow \mathbb{R}$  that satisfies

- ▶  $L$ -smoothness:  $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ ;
- ▶ Strong convexity:  $(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) \geq m\|\mathbf{x} - \mathbf{y}\|^2$  for some  $m > 0$ .

Consider a Markov chain  $\mathbf{x}_{I+1} = \mathbf{x}_I + \frac{\eta}{2} \cdot \nabla_{\mathbf{x}_I} \log \hat{p}(\mathbf{x}_I) + \sqrt{\eta} \cdot \boldsymbol{\epsilon}_I$ , where  $\boldsymbol{\epsilon}_I \sim \mathcal{N}(0, \mathbf{I})$ . Then, for any  $\eta < \frac{2}{L}$

- ▶ The Markov chain has a unique stationary distribution  $\pi_\eta$ .
- ▶  $W_2(\pi_\eta, p) \leq C\eta$ , and as  $\eta \rightarrow 0$  we have  $\pi_\eta \xrightarrow{d} p$ .

# Langevin Dynamics

## Theorem (Informal)

Let  $\mathbf{x}_0$  be a random vector. Under mild regularity conditions, samples from the following dynamics will eventually follow  $p_\theta(\mathbf{x})$  (for sufficiently small  $\eta$  and large  $I$ ):

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

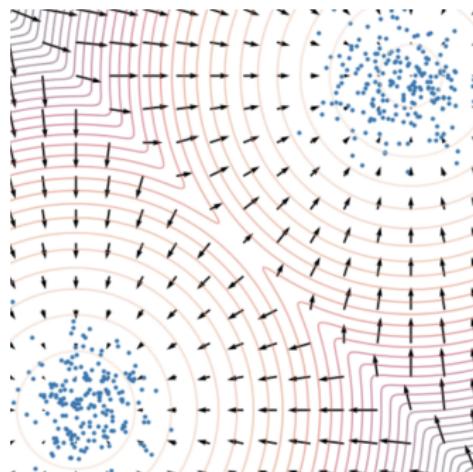
# Langevin Dynamics

## Theorem (Informal)

Let  $\mathbf{x}_0$  be a random vector. Under mild regularity conditions, samples from the following dynamics will eventually follow  $p_\theta(\mathbf{x})$  (for sufficiently small  $\eta$  and large  $I$ ):

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

- ▶ What if  $\boldsymbol{\epsilon}_I = \mathbf{0}$ ?
- ▶ The density  $p_\theta(\mathbf{x})$  is the **stationary** distribution of the Markov chain.
- ▶ The gradient is taken with respect to  $\mathbf{x}$ , not  $\theta$ .
- ▶  $\nabla_{\mathbf{x}} \log p_\theta(\mathbf{x})$  defines a vector field.



# Outline

## 1. Evaluation of Likelihood-Free Models

Frechet Inception Distance (FID)

Precision-Recall

CLIP Score

Human Evaluation

## 2. Langevin Dynamics

## 3. Score Matching

## 4. Denoising Score Matching

# Score Matching

## Score Function

$$s_{\theta}(x) = \nabla_x \log p_{\theta}(x)$$

# Score Matching

## Score Function

$$\mathbf{s}_\theta(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_\theta(\mathbf{x})$$

## Langevin Dynamics

If we know the score function  $\mathbf{s}_\theta(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_\theta(\mathbf{x})$ , we can generate samples from the model using Langevin dynamics:

$$\mathbf{x}_{I+1} = \mathbf{x}_I + \frac{\eta}{2} \cdot \nabla_{\mathbf{x}_I} \log p_\theta(\mathbf{x}_I) + \sqrt{\eta} \cdot \epsilon_I = \mathbf{x}_I + \frac{\eta}{2} \cdot \mathbf{s}_\theta(\mathbf{x}_I) + \sqrt{\eta} \cdot \epsilon_I.$$

# Score Matching

## Score Function

$$\mathbf{s}_\theta(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_\theta(\mathbf{x})$$

## Langevin Dynamics

If we know the score function  $\mathbf{s}_\theta(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_\theta(\mathbf{x})$ , we can generate samples from the model using Langevin dynamics:

$$\mathbf{x}_{I+1} = \mathbf{x}_I + \frac{\eta}{2} \cdot \nabla_{\mathbf{x}_I} \log p_\theta(\mathbf{x}_I) + \sqrt{\eta} \cdot \epsilon_I = \mathbf{x}_I + \frac{\eta}{2} \cdot \mathbf{s}_\theta(\mathbf{x}_I) + \sqrt{\eta} \cdot \epsilon_I.$$

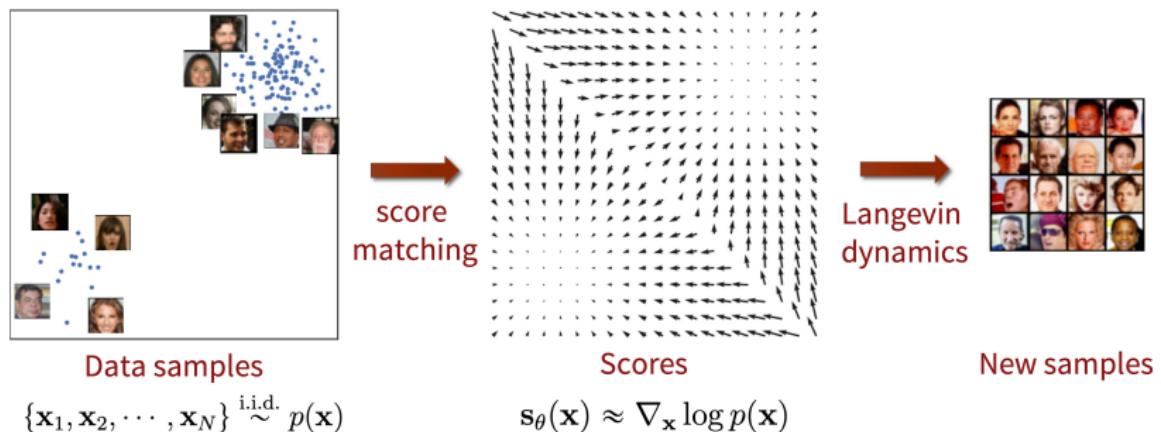
## Fisher Divergence

$$\begin{aligned} D_F(p_{\text{data}}, p_\theta) &= \frac{1}{2} \mathbb{E}_\pi \left\| \nabla_{\mathbf{x}} \log p_\theta(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x}) \right\|_2^2 = \\ &= \frac{1}{2} \mathbb{E}_\pi \left\| \mathbf{s}_\theta(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x}) \right\|_2^2 \rightarrow \min_{\theta} \end{aligned}$$

# Score Matching

## Fisher Divergence

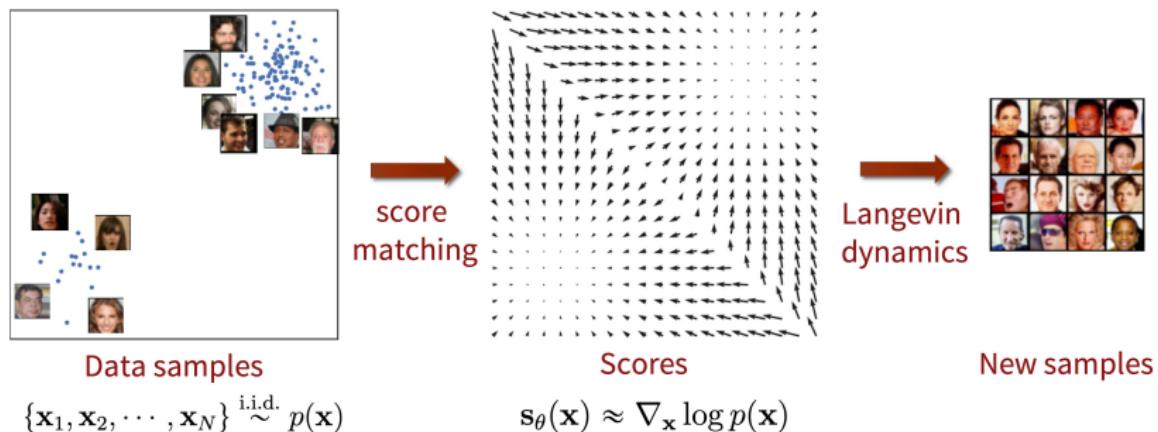
$$D_F(p_{\text{data}}, p_{\theta}) = \frac{1}{2} \mathbb{E}_{\pi} \| \mathbf{s}_{\theta}(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x}) \|_2^2 \rightarrow \min_{\theta}$$



# Score Matching

## Fisher Divergence

$$D_F(p_{\text{data}}, p_{\theta}) = \frac{1}{2} \mathbb{E}_{\pi} \| \mathbf{s}_{\theta}(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x}) \|_2^2 \rightarrow \min_{\theta}$$



**Problem:** We don't know  $\nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x})$ .

# Outline

## 1. Evaluation of Likelihood-Free Models

Frechet Inception Distance (FID)

Precision-Recall

CLIP Score

Human Evaluation

## 2. Langevin Dynamics

## 3. Score Matching

## 4. Denoising Score Matching

## Denoising Score Matching

Let us perturb the original data  $\mathbf{x} \sim p_{\text{data}}(\mathbf{x})$  with Gaussian noise:

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

## Denoising Score Matching

Let us perturb the original data  $\mathbf{x} \sim p_{\text{data}}(\mathbf{x})$  with Gaussian noise:

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

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

## Denoising Score Matching

Let us perturb the original data  $\mathbf{x} \sim p_{\text{data}}(\mathbf{x})$  with Gaussian noise:

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

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

### Assumption

The solution to

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

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

# Denoising Score Matching

Let us perturb the original data  $\mathbf{x} \sim p_{\text{data}}(\mathbf{x})$  with Gaussian noise:

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

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

## Assumption

The solution to

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

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

- ▶ The score function of the noised data nearly matches the score function of the original data.
- ▶ The score function  $\mathbf{s}_{\theta, \sigma}(\mathbf{x}_\sigma)$  is parameterized by  $\sigma$ .
- ▶ **Note:** We don't know  $q(\mathbf{x}_\sigma)$ , just as we don't know  $p_{\text{data}}(\mathbf{x})$ .

# Denoising Score Matching

## Theorem

Under mild regularity conditions, the following holds:

$$\begin{aligned}\mathbb{E}_{q(\mathbf{x}_\sigma)} \left\| \mathbf{s}_{\theta, \sigma}(\mathbf{x}_\sigma) - \nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma) \right\|_2^2 &= \\ &= \mathbb{E}_{p_{\text{data}}(\mathbf{x})} \mathbb{E}_{q(\mathbf{x}_\sigma | \mathbf{x})} \left\| \mathbf{s}_{\theta, \sigma}(\mathbf{x}_\sigma) - \nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma | \mathbf{x}) \right\|_2^2 + \text{const}(\theta)\end{aligned}$$

# Denoising Score Matching

## Theorem

Under mild regularity conditions, the following holds:

$$\begin{aligned}\mathbb{E}_{q(\mathbf{x}_\sigma)} \left\| \mathbf{s}_{\theta, \sigma}(\mathbf{x}_\sigma) - \nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma) \right\|_2^2 &= \\ &= \mathbb{E}_{p_{\text{data}}(\mathbf{x})} \mathbb{E}_{q(\mathbf{x}_\sigma | \mathbf{x})} \left\| \mathbf{s}_{\theta, \sigma}(\mathbf{x}_\sigma) - \nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma | \mathbf{x}) \right\|_2^2 + \text{const}(\theta)\end{aligned}$$

## Gradient of the Noise Kernel

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

# Denoising Score Matching

## Theorem

Under mild regularity conditions, the following holds:

$$\begin{aligned}\mathbb{E}_{q(\mathbf{x}_\sigma)} \left\| \mathbf{s}_{\theta, \sigma}(\mathbf{x}_\sigma) - \nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma) \right\|_2^2 &= \\ &= \mathbb{E}_{p_{\text{data}}(\mathbf{x})} \mathbb{E}_{q(\mathbf{x}_\sigma | \mathbf{x})} \left\| \mathbf{s}_{\theta, \sigma}(\mathbf{x}_\sigma) - \nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma | \mathbf{x}) \right\|_2^2 + \text{const}(\theta)\end{aligned}$$

## Gradient of the Noise Kernel

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

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

# Denoising Score Matching

## Theorem

Under mild regularity conditions, the following holds:

$$\begin{aligned}\mathbb{E}_{q(\mathbf{x}_\sigma)} \left\| \mathbf{s}_{\theta, \sigma}(\mathbf{x}_\sigma) - \nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma) \right\|_2^2 &= \\ &= \mathbb{E}_{p_{\text{data}}(\mathbf{x})} \mathbb{E}_{q(\mathbf{x}_\sigma | \mathbf{x})} \left\| \mathbf{s}_{\theta, \sigma}(\mathbf{x}_\sigma) - \nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma | \mathbf{x}) \right\|_2^2 + \text{const}(\theta)\end{aligned}$$

## 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 right-hand side doesn't require computing  $\nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma)$  or even  $\nabla_{\mathbf{x}_\sigma} \log p_{\text{data}}(\mathbf{x}_\sigma)$ .
- ▶  $\mathbf{s}_{\theta, \sigma}(\mathbf{x}_\sigma)$  is trained to **denoise** the noised samples  $\mathbf{x}_\sigma$ .

# Denoising Score Matching

Initial objective:

$$\mathbb{E}_{p_{\text{data}}(\mathbf{x})} \left\| \mathbf{s}_\theta(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x}) \right\|_2^2 \rightarrow \min_{\theta}$$

# Denoising Score Matching

Initial objective:

$$\mathbb{E}_{p_{\text{data}}(\mathbf{x})} \|\mathbf{s}_\theta(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x})\|_2^2 \rightarrow \min_{\theta}$$

Noised objective:

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

# Denoising Score Matching

Initial objective:

$$\mathbb{E}_{p_{\text{data}}(\mathbf{x})} \|\mathbf{s}_\theta(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x})\|_2^2 \rightarrow \min_{\theta}$$

Noised objective:

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

This is equivalent to a denoising task:

$$\mathbb{E}_{p_{\text{data}}(\mathbf{x})} \mathbb{E}_{q(\mathbf{x}_\sigma|\mathbf{x})} \|\mathbf{s}_{\theta,\sigma}(\mathbf{x}_\sigma) - \nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma|\mathbf{x})\|_2^2 \rightarrow \min_{\theta}$$

# Denoising Score Matching

Initial objective:

$$\mathbb{E}_{p_{\text{data}}(\mathbf{x})} \|\mathbf{s}_\theta(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x})\|_2^2 \rightarrow \min_{\theta}$$

Noised objective:

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

This is equivalent to a denoising task:

$$\mathbb{E}_{p_{\text{data}}(\mathbf{x})} \mathbb{E}_{q(\mathbf{x}_\sigma|\mathbf{x})} \|\mathbf{s}_{\theta,\sigma}(\mathbf{x}_\sigma) - \nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma|\mathbf{x})\|_2^2 \rightarrow \min_{\theta}$$

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

## Langevin Dynamics

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

## Summary

- ▶ Frechet Inception Distance is the most popular metric for evaluating implicit generative models.
- ▶ Precision-recall allows for choosing a model that balances sample quality and diversity.
- ▶ The CLIP score is widely used to measure text-to-image alignment.
- ▶ The gold standard for evaluating generated image quality is human assessment.
- ▶ Langevin dynamics enable sampling from generative models using gradients of the log-likelihood.
- ▶ Score matching proposes minimizing Fisher divergence to estimate the score function.
- ▶ Denoising score matching optimizes Fisher divergence on noisy data, making it estimable with samples.