

# Deep Generative Models

## Lecture 9

Roman Isachenko



2024, Summer

## Recap of previous lecture

### Wasserstein distance

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

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

### Theorem (Kantorovich-Rubinstein duality)

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

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

## Recap of previous lecture

### WGAN objective

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

- ▶ Function  $f$  in WGAN is usually called *critic*.
- ▶ If parameters  $\phi$  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 distributions in  $\mathcal{X}$ , a compact metric space. Let  $\gamma$  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  $\hat{\mathbf{x}}_t = t \cdot \mathbf{y} + (1 - t) \cdot \mathbf{z}$  with  $t \in [0, 1]$  are uniformly sampled along straight lines between pairs of points:  $\mathbf{y} \sim \pi(\mathbf{x})$  and  $\mathbf{z} \sim p(\mathbf{x}|\theta)$ .

## Recap of previous lecture

### f-divergence minimization

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

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

### Variational divergence estimation

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

### Fenchel conjugate

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

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

# Outline

## 1. Evaluation of likelihood-free models

Frechet Inception Distance (FID)

Maximum Mean Discrepancy (MMD)

Precision-Recall

## 2. Langevin dynamic

## 3. Score matching

# Outline

## 1. Evaluation of likelihood-free models

Frechet Inception Distance (FID)

Maximum Mean Discrepancy (MMD)

Precision-Recall

## 2. Langevin dynamic

## 3. Score matching

# 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).

# Outline

## 1. Evaluation of likelihood-free models

Frechet Inception Distance (FID)

Maximum Mean Discrepancy (MMD)

Precision-Recall

## 2. Langevin dynamic

## 3. Score matching

# Frechet Inception Distance (FID)

Wasserstein metric

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

Theorem

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

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

## Frechet Inception Distance

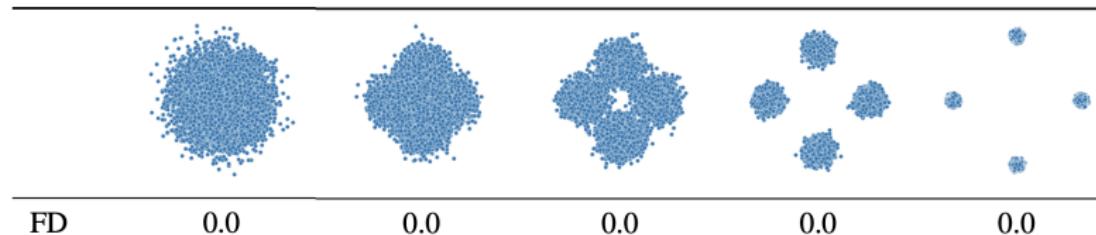
$$\text{FID}(\pi, p) = W_2^2(\pi, p)$$

- ▶ 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  $p(\mathbf{x}|\theta)$ .

## Frechet Inception Distance (FID)

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

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



# Outline

## 1. Evaluation of likelihood-free models

Frechet Inception Distance (FID)

Maximum Mean Discrepancy (MMD)

Precision-Recall

## 2. Langevin dynamic

## 3. Score matching

# Maximum Mean Discrepancy (MMD)

## Theorem

$\pi(\mathbf{x}) = p(\mathbf{y})$  if and only if  $\mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}) = \mathbb{E}_{p(\mathbf{y})} f(\mathbf{y})$  for any bounded and continuous  $f$ .

$$\text{MMD}(\pi, p) = \sup_f [\mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} f(\mathbf{x})]$$

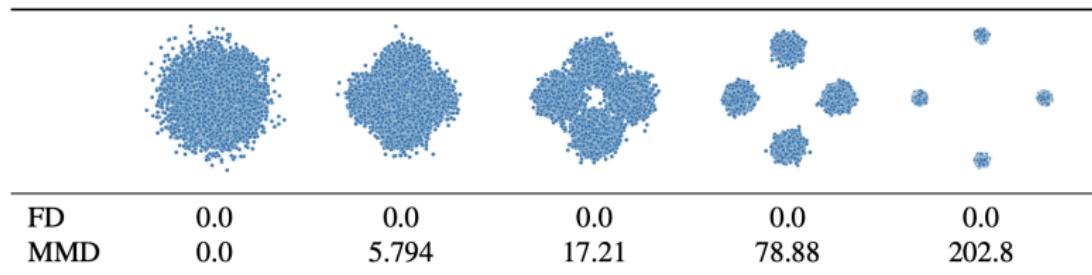
## Theorem (Reproducing Kernel Hilbert Space)

$$\text{MMD}^2(\pi, p) = \mathbb{E}_{\substack{\mathbf{x} \sim \pi(\mathbf{x}) \\ \mathbf{x}' \sim \pi(\mathbf{x})}} k(\mathbf{x}, \mathbf{x}') + \mathbb{E}_{\substack{\mathbf{y} \sim p(\mathbf{y}) \\ \mathbf{y}' \sim p(\mathbf{y})}} k(\mathbf{y}, \mathbf{y}') - 2 \mathbb{E}_{\substack{\mathbf{x} \sim \pi(\mathbf{x}) \\ \mathbf{y} \sim p(\mathbf{y})}} k(\mathbf{x}, \mathbf{y})$$

- ▶  $k(\mathbf{x}, \mathbf{y})$  is a positive definite, symmetric kernel function (for example  $k(\mathbf{x}, \mathbf{y}) = \frac{\exp(-\|\mathbf{x}-\mathbf{y}\|^2)}{\sigma^2}$ );
- ▶  $f(\mathbf{x}) = \langle f, \phi(\mathbf{x}) \rangle_{\mathcal{H}}$ ;
- ▶  $\phi(\mathbf{x}) = k(\cdot, \mathbf{x})$ .

# Maximum Mean Discrepancy (MMD)

$$\text{MMD}^2(\pi, p) = \mathbb{E}_{\substack{\mathbf{x} \sim \pi(\mathbf{x}) \\ \mathbf{x}' \sim \pi(\mathbf{x})}} k(\mathbf{x}, \mathbf{x}') + \mathbb{E}_{\substack{\mathbf{y} \sim p(\mathbf{y}) \\ \mathbf{y}' \sim p(\mathbf{y})}} k(\mathbf{y}, \mathbf{y}') - 2\mathbb{E}_{\substack{\mathbf{x} \sim \pi(\mathbf{x}) \\ \mathbf{y} \sim p(\mathbf{y})}} k(\mathbf{x}, \mathbf{y})$$



- ▶ Needs less sample size for evaluation.
- ▶ High dependence on the pretrained classification model.
- ▶ Works with any distribution!

# Outline

## 1. Evaluation of likelihood-free models

Frechet Inception Distance (FID)

Maximum Mean Discrepancy (MMD)

Precision-Recall

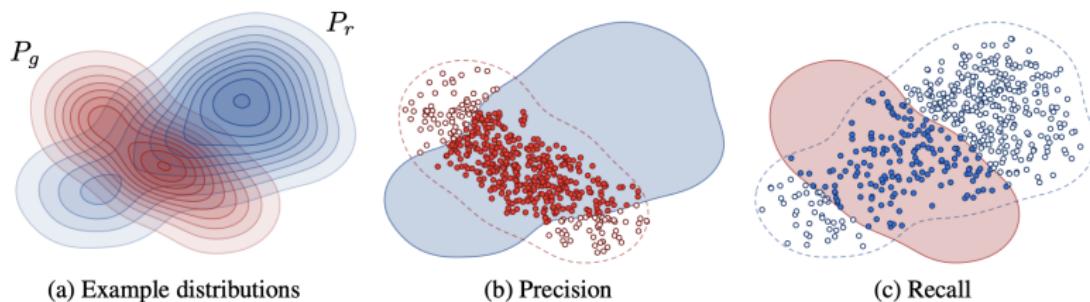
## 2. Langevin dynamic

## 3. Score matching

# 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.



- ▶ **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

- ▶  $\mathcal{S}_\pi = \{\mathbf{x}_i\}_{i=1}^n \sim \pi(\mathbf{x})$  – 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):

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

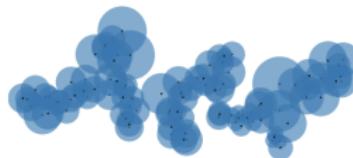
Define binary function:

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

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

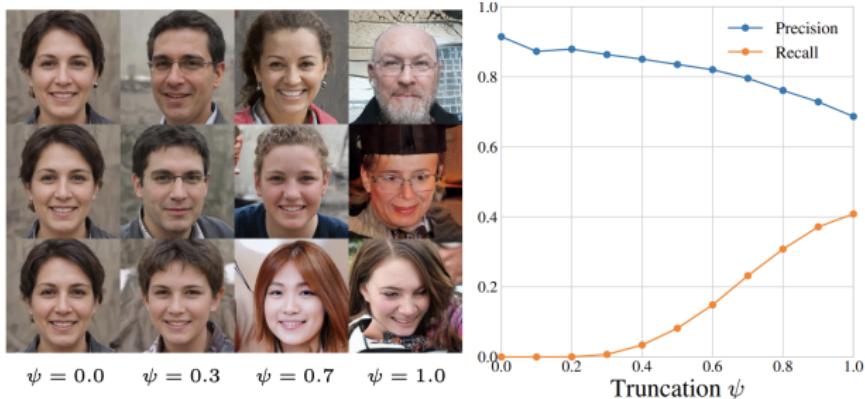
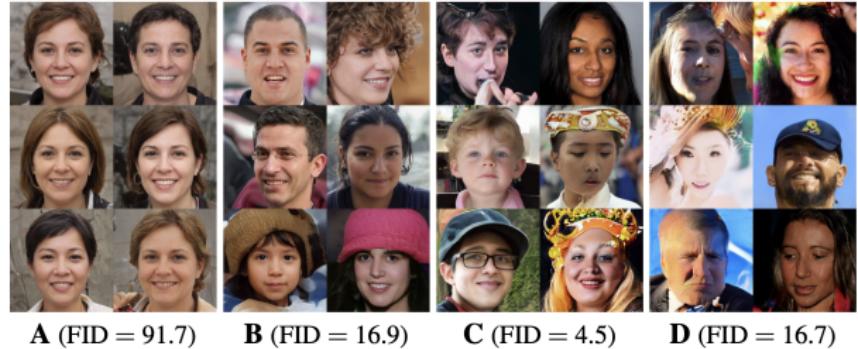
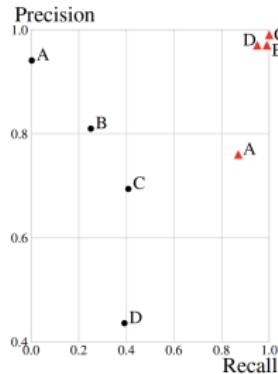


(a) True manifold



(b) Approx. manifold

# Precision-Recall



## 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  $\psi$  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

1. Evaluation of likelihood-free models
  - Frechet Inception Distance (FID)
  - Maximum Mean Discrepancy (MMD)
  - Precision-Recall
2. Langevin dynamic
3. Score matching

# Langevin dynamic

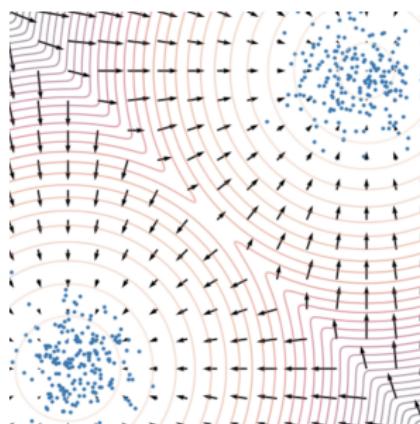
## Statement

Let  $\mathbf{x}_0$  be a random vector. Then samples from the following dynamics

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

will come from  $p(\mathbf{x} | \theta)$  (under mild regularity conditions, for small enough  $\eta$  and large enough  $l$ ).

- ▶ Here we assume that we already have some generative model  $p(\mathbf{x} | \theta)$ .
- ▶ The density  $p(\mathbf{x} | \theta)$  is a **stationary** distribution for the Markov chain.
- ▶ What do we get if  $\epsilon = \mathbf{0}$ ?



## Energy-based models

### Langevin dynamic

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

We are able to sample from the model using Langevin dynamics if we have  $\nabla_{\mathbf{x}} \log p(\mathbf{x} | \boldsymbol{\theta})$ .

### 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.

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

The gradient of the normalized density equals to the gradient of the unnormalized density.

# Outline

1. Evaluation of likelihood-free models
  - Frechet Inception Distance (FID)
  - Maximum Mean Discrepancy (MMD)
  - Precision-Recall
2. Langevin dynamic
3. Score matching

# Score matching

## Score function

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

## Langevin dynamic

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

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

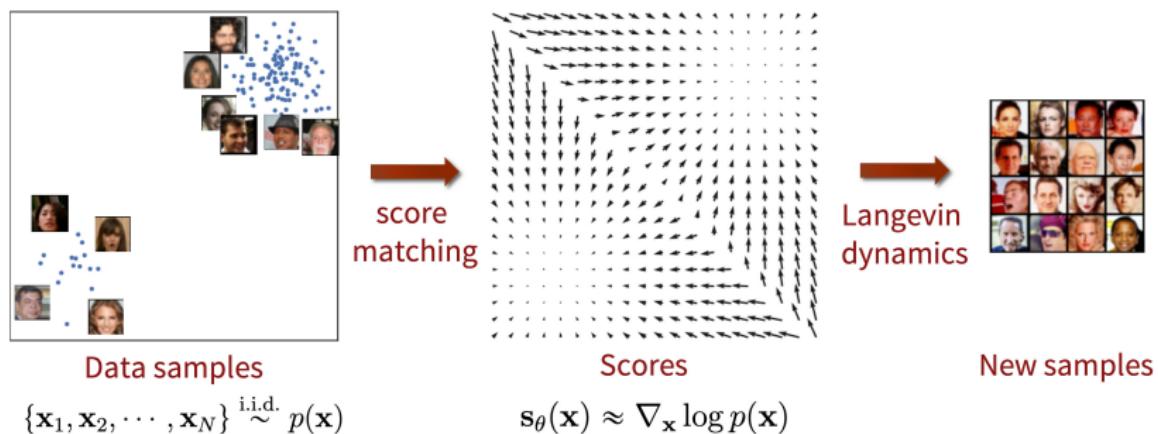
## Fisher divergence

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

# Score matching

## Fisher divergence

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



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

## Summary

- ▶ We need a measure of quality for the implicit models (like GANs).
- ▶ Frechet Inception Distance is the most popular metric for the implicit models evaluation.
- ▶ Maximum Mean Discrepancy tries to fix some of the FID drawbacks.
- ▶ 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.
- ▶ 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.