

# Deep Generative Models

## Lecture 11

Roman Isachenko

Moscow Institute of Physics and Technology

2022

## 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

$$W(\pi || p) = \underbrace{\mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} f(\mathbf{x})}_{\text{original critic loss}} + \lambda \underbrace{\mathbb{E}_{U[0,1]} \left[ (\|\nabla_{\hat{\mathbf{x}}} f(\hat{\mathbf{x}})\|_2 - 1)^2 \right]}_{\text{gradient penalty}}.$$

Samples  $\hat{\mathbf{x}}_t = t\mathbf{y} + (1-t)\mathbf{z}$  with  $t \in [0, 1]$  are uniformly sampled along straight lines between pairs of points:  $\mathbf{y}$  from the data distribution  $\pi(\mathbf{x})$  and  $\mathbf{z}$  from the generator distribution  $p(\mathbf{x}|\theta)$ .

## Recap of previous lecture

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

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

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

### Critic spectral norm

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

### Spectral Normalization GAN

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

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

## 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.

## Recap of previous lecture

How to evaluate likelihood-free models?

$p(y|x)$  – pretrained image classification model (e.g. ImageNet classifier).

What do we want from samples?

- ▶ Sharpness



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

- ▶ Diversity



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

# Outline

## 1. Evaluation of likelihood-free models

Inception score

Frechet Inception Distance

Precision-Recall

## 2. Discrete VAE latent representations

# Outline

## 1. Evaluation of likelihood-free models

Inception score

Frechet Inception Distance

Precision-Recall

## 2. Discrete VAE latent representations

# Outline

## 1. Evaluation of likelihood-free models

Inception score

Frechet Inception Distance

Precision-Recall

## 2. Discrete VAE latent representations

# Evaluation of likelihood-free models

What do we want from samples?

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

Inception Score

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

# Outline

## 1. Evaluation of likelihood-free models

Inception score

Frechet Inception Distance

Precision-Recall

## 2. Discrete VAE latent representations

# Evaluation of likelihood-free models

## Theorem (informal)

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

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

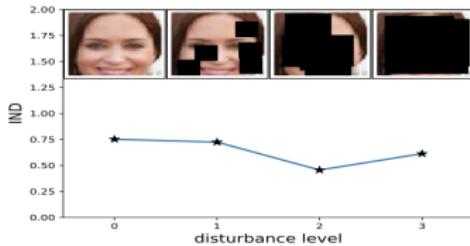
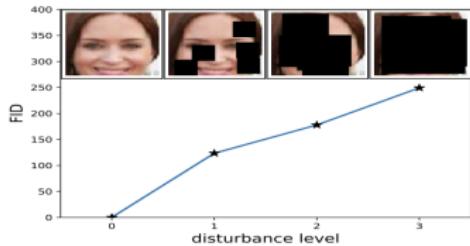
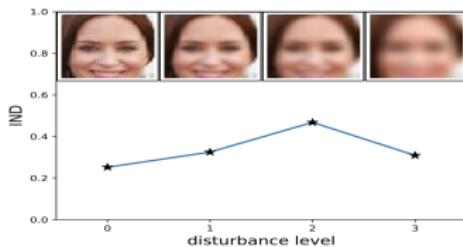
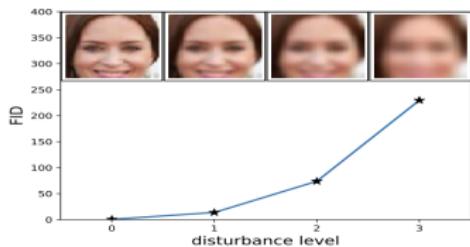
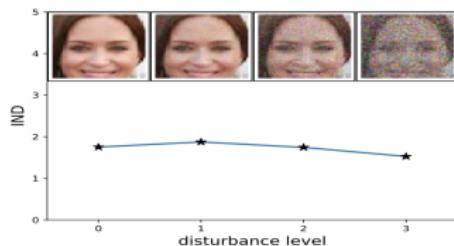
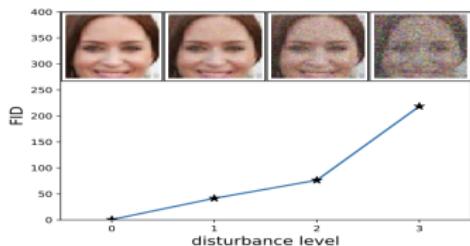
This is intractable to calculate all moments.

## Frechet Inception Distance

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

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

# Evaluation of likelihood-free models



# Limitations

## Inception Score

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

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

## Frechet Inception Distance

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

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

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

---

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

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

# Outline

## 1. Evaluation of likelihood-free models

Inception score

Frechet Inception Distance

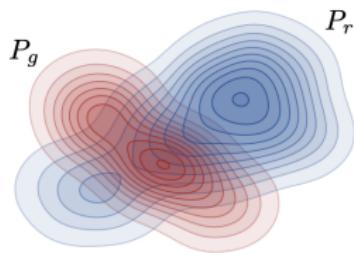
Precision-Recall

## 2. Discrete VAE latent representations

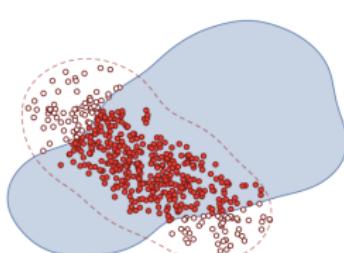
# Precision-Recall for Generative Models

What do we want from samples

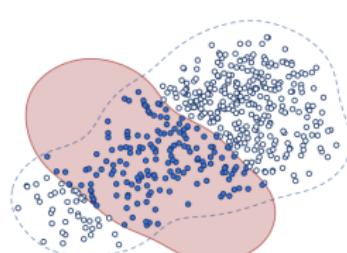
- ▶ **Sharpness:** generated samples should be of high quality.
- ▶ **Diversity:** their variation should match that observed in the training set.



(a) Example distributions



(b) Precision



(c) Recall

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

## Precision-Recall for generative models

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

Embed samples using pretrained classifier network (as previously):

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

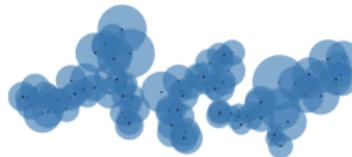
Define binary function:

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

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

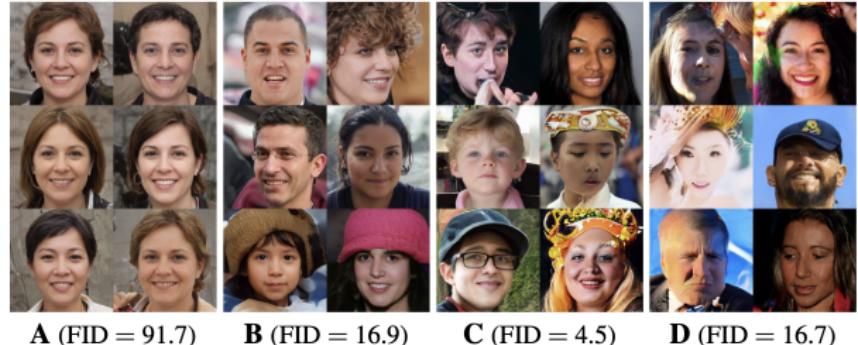
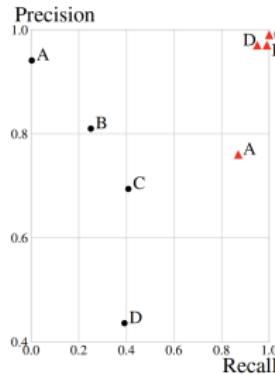


(a) True manifold

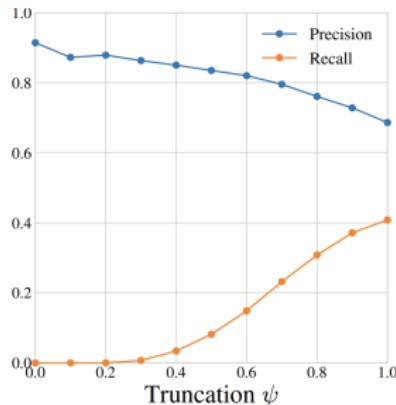


(b) Approx. manifold

# Precision-Recall for generative models



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



## 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

Inception score

Frechet Inception Distance

Precision-Recall

## 2. Discrete VAE latent representations

# Discrete VAE latents

## Motivation

- ▶ Previous VAE models had **continuous** latent variables  $\mathbf{z}$ .
- ▶ **Discrete** representations  $\mathbf{z}$  are potentially a more natural fit for many of the modalities.
- ▶ Powerful autoregressive models (like PixelCNN) have been developed for modelling distributions over discrete variables.
- ▶ All cool transformer-like models work with discrete tokens.

## ELBO

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \phi)} \log p(\mathbf{x}|\mathbf{z}, \theta) - KL(q(\mathbf{z}|\mathbf{x}, \phi) || p(\mathbf{z})) \rightarrow \max_{\phi, \theta} .$$

- ▶ Reparametrization trick to get unbiased gradients.
- ▶ Normal assumptions for  $q(\mathbf{z}|\mathbf{x}, \phi)$  and  $p(\mathbf{z})$  to compute KL analytically.

# Discrete VAE latents

## Assumptions

- ▶ Define dictionary (word book) space  $\{\mathbf{e}_k\}_{k=1}^K$ , where  $\mathbf{e}_k \in \mathbb{R}^C$ ,  $K$  is the size of the dictionary.
- ▶ Let  $c \sim \text{Categorical}(\boldsymbol{\pi})$ , where
$$\boldsymbol{\pi} = (\pi_1, \dots, \pi_K), \quad \pi_k = P(c = k), \quad \sum_{k=1}^K \pi_k = 1.$$
- ▶ Let VAE model has discrete latent representation  $c$  with prior  $p(c) = \text{Uniform}\{1, \dots, K\}$ .

## How it should work?

- ▶ Our variational posterior  $q(c|\mathbf{x}, \phi) = \text{Categorical}(\boldsymbol{\pi}(\mathbf{x}, \phi))$  (encoder) outputs discrete probabilities vector.
- ▶ We sample  $c^*$  from  $q(c|\mathbf{x}, \phi)$  (reparametrization trick analogue).
- ▶ Our generative distribution  $p(\mathbf{x}|\mathbf{e}_{c^*}, \theta)$  (decoder).

# Discrete VAE latents

## ELBO

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_{q(c|\mathbf{x}, \phi)} \log p(\mathbf{x}|c, \theta) - KL(q(c|\mathbf{x}, \phi)||p(c)) \rightarrow \max_{\phi, \theta} .$$

### KL term

$$\begin{aligned} KL(q(c|\mathbf{x}, \phi)||p(c)) &= \sum_{k=1}^K q(k|\mathbf{x}, \phi) \log \frac{q(k|\mathbf{x}, \phi)}{p(k)} = \\ &= \sum_{k=1}^K q(k|\mathbf{x}, \phi) \log q(k|\mathbf{x}, \phi) - \sum_{k=1}^K q(k|\mathbf{x}, \phi) \log p(k) = \\ &= -H(q(c|\mathbf{x}, \phi)) + \log K. \end{aligned}$$

- ▶ Is it possible to make reparametrization trick? (we sample from discrete distribution now!).
- ▶ Entropy term should be estimated.

## Summary

- ▶ Inception Score and Frechet Inception Distance are the common metrics for GAN evaluation, but both of them have drawbacks.
- ▶ Precision-recall allows to select model that compromises the sample quality and the sample diversity.
- ▶ Truncation tricks help to select model with compromised samples: diverse and sharp.
- ▶ Discrete VAE latents is a natural idea, but we have to avoid non-differentiable sampling operation.