Deep Generative Models

Lecture 9

Roman Isachenko

AI Masters

2024, Spring

Recap of previous lecture

Main problems of standard GAN

- Vanishing gradients (solution: non-saturating GAN);
- Mode collapse (caused by Jensen-Shannon divergence).

Standard GAN

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

Informal theoretical results

The real images distribution $\pi(\mathbf{x})$ and the generated images distribution $p(\mathbf{x}|\boldsymbol{\theta})$ are low-dimensional and have disjoint supports. In this case

$$KL(\pi || p) = KL(p || \pi) = \infty, \quad JSD(\pi || p) = \log 2.$$

Goodfellow I. J. et al. Generative Adversarial Networks, 2014
Arjovsky M., Bottou L. Towards Principled Methods for Training Generative Adversarial Networks, 2017

Recap of previous lecture

Wasserstein distance

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

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

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 ($f : \mathcal{X} \to \mathbb{R}$).

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

Recap of previous lecture

WGAN objective

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

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

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

"Weight clipping is a clearly terrible way to enforce a Lipschitz constraint"

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

Recap of previous lecture

Theorem

Let $\pi(\mathbf{x})$ and $p(\mathbf{x})$ be two distribution in $\mathcal{X}$, a compact metric space. Let γ 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 $\mathbf{z} \sim p(\mathbf{x}|\boldsymbol{\theta})$.

Arjovsky M., Chintala S., Bottou L. Wasserstein GAN, 2017
Gulrajani I. et al. Improved Training of Wasserstein GANs, 2017

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

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

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.

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

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 take some pretrained image classification model to get the conditional label distribution $p(y|\mathbf{x})$ (e.g. ImageNet classifier).

What do we want from samples?

- **Sharpness**

  Low sharpness High sharpness

  The conditional distribution $p(y|\mathbf{x})$ should have low entropy (each image $\mathbf{x}$ should have distinctly recognizable object).

- **Diversity**

  Low diversity High diversity

  The marginal distribution $p(y) = \int p(y|\mathbf{x})p(\mathbf{x})d\mathbf{x}$ should have high entropy (there should be as many classes generated as possible).

---

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

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

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

Theorem

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

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

Frechet Inception Distance

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

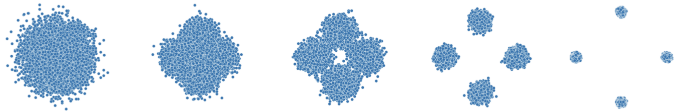
- Representations are the outputs of the intermediate layer from the pretrained classification model.
- $\boldsymbol{\mu}_\pi$, $\boldsymbol{\Sigma}_\pi$ and $\boldsymbol{\mu}_p$, $\boldsymbol{\Sigma}_p$ are the statistics of the feature representations for the samples from $\pi(\mathbf{x})$ and $p(\mathbf{x}|\boldsymbol{\theta})$.

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

Frechet Inception Distance (FID)

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

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



FD	0.0	0.0	0.0	0.0	0.0

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

Outline

1. 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.

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

Theorem (Reproducing Kernel Hilbert Space)

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

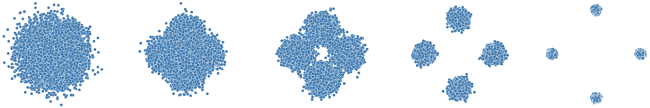
Here

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

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

Maximum Mean Discrepancy (MMD)

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



FD	0.0	0.0	0.0	0.0	0.0
MMD	0.0	5.794	17.21	78.88	202.8

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

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

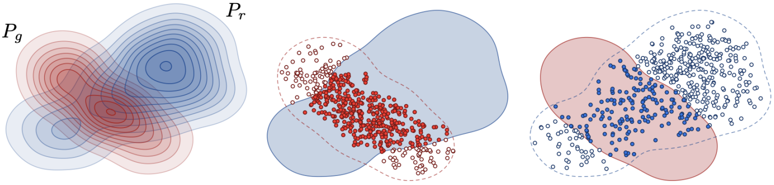
Outline

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.



(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.

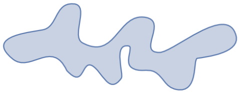
Embed samples using pretrained classifier network (as previously):

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

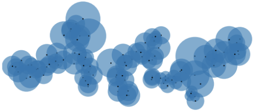
Define binary function:

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

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



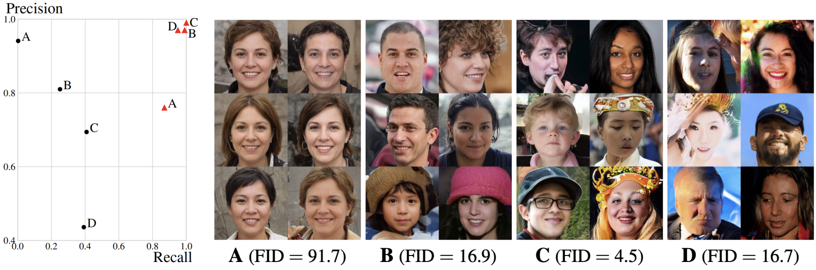
(a) True manifold

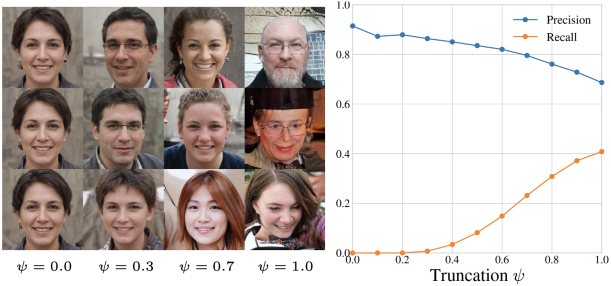


(b) Approx. manifold

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

Precision-Recall





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

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 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

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

Langevin dynamic

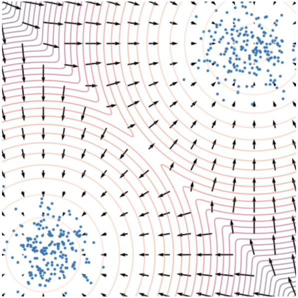
Statement

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

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

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

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



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

Energy-based models

Langevin dynamic

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

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

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

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.

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

Fisher divergence

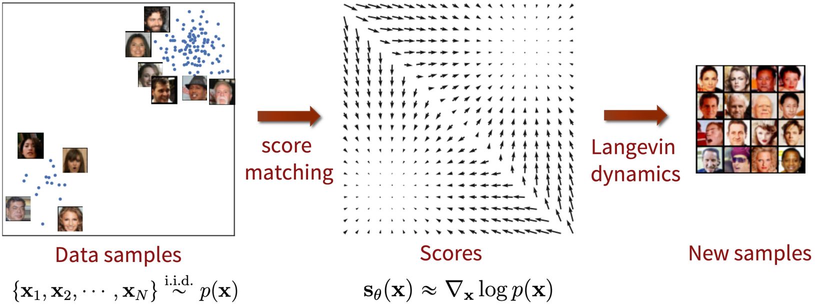
$$D_F(\pi, p) = \frac{1}{2}\mathbb{E}_\pi \big\| \nabla_{\mathbf{x}} \log p(\mathbf{x}|\boldsymbol{\theta}) - \nabla_{\mathbf{x}} \log \pi(\mathbf{x}) \big\|_2^2 =$$
$$= \frac{1}{2}\mathbb{E}_\pi \big\| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_l) - \nabla_{\mathbf{x}} \log \pi(\mathbf{x}) \big\|_2^2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

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

Score matching

Fisher divergence

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



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

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 dynamic 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.