

# Deep Generative Models

## Lecture 13

Roman Isachenko



Spring, 2022

## Recap of previous lecture

Continuous dynamic

$$\frac{d\mathbf{z}(t)}{dt} = f(\mathbf{z}(t), \theta).$$

Forward pass

$$\mathbf{z}(t_1) = \int_{t_0}^{t_1} f(\mathbf{z}(t), \theta) dt + \mathbf{z}_0 \quad \Rightarrow \quad \text{ODE Solver}$$

Backward pass

$$\left. \begin{aligned} \frac{\partial L}{\partial \theta(t_0)} &= \mathbf{a}_\theta(t_0) = - \int_{t_1}^{t_0} \mathbf{a}_z(t)^\top \frac{\partial f(\mathbf{z}(t), \theta)}{\partial \theta(t)} dt + 0 \\ \frac{\partial L}{\partial \mathbf{z}(t_0)} &= \mathbf{a}_z(t_0) = - \int_{t_1}^{t_0} \mathbf{a}_z(t)^\top \frac{\partial f(\mathbf{z}(t), \theta)}{\partial \mathbf{z}(t)} dt + \frac{\partial L}{\partial \mathbf{z}(t_1)} \\ \mathbf{z}(t_0) &= - \int_{t_1}^{t_0} f(\mathbf{z}(t), \theta) dt + \mathbf{z}_1. \end{aligned} \right\} \Rightarrow \text{ODE Solver}$$

## Recap of previous lecture

### Continuous normalizing flows

$$\frac{d \log p(\mathbf{z}(t), t)}{dt} = -\text{tr} \left( \frac{\partial f(\mathbf{z}(t), \boldsymbol{\theta})}{\partial \mathbf{z}(t)} \right).$$

Forward transform + log-density

$$\begin{bmatrix} \mathbf{x} \\ \log p(\mathbf{x}|\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \mathbf{z} \\ \log p(\mathbf{z}) \end{bmatrix} + \int_{t_0}^{t_1} \begin{bmatrix} f(\mathbf{z}(t), \boldsymbol{\theta}) \\ -\text{tr} \left( \frac{\partial f(\mathbf{z}(t), \boldsymbol{\theta})}{\partial \mathbf{z}(t)} \right) \end{bmatrix} dt.$$

Hutchinson's trace estimator

$$\begin{aligned} \log p(\mathbf{z}(t_1)) &= \log p(\mathbf{z}(t_0)) - \int_{t_0}^{t_1} \text{tr} \left( \frac{\partial f(\mathbf{z}(t), \boldsymbol{\theta})}{\partial \mathbf{z}(t)} \right) dt = \\ &= \log p(\mathbf{z}(t_0)) - \mathbb{E}_{p(\epsilon)} \int_{t_0}^{t_1} \left[ \epsilon^T \frac{\partial f}{\partial \mathbf{z}} \epsilon \right] dt. \end{aligned}$$

## Recap of previous lecture

### SDE basics

Let define stochastic process  $\mathbf{x}(t)$  with initial condition  $\mathbf{x}(0) \sim p_0(\mathbf{x})$ :

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w},$$

where  $\mathbf{w}(t)$  is the standard Wiener process (Brownian motion)

$$\mathbf{w}(t) - \mathbf{w}(s) \sim \mathcal{N}(0, (t-s)\mathbf{I}), \quad d\mathbf{w} = \epsilon \cdot \sqrt{dt}, \text{ where } \epsilon \sim \mathcal{N}(0, \mathbf{I}).$$

### Langevin dynamics

Let  $\mathbf{x}_0$  be a random vector. Then under mild regularity conditions for small enough  $\eta$  samples from the following dynamics

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

will come from  $p(\mathbf{x} | \theta)$ .

The density  $p(\mathbf{x} | \theta)$  is a **stationary** distribution for the Langevin SDE.

# Outline

1. Score matching
2. Diffusion models

# Outline

1. Score matching
2. Diffusion models

## Score matching

We could sample from the model if we have  $\nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta)$ .

## Fisher divergence

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

## Score function

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

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

## Theorem

Under some regularity conditions, it holds

$$\frac{1}{2} \mathbb{E}_{\pi} \| \mathbf{s}(\mathbf{x}, \theta) - \nabla_{\mathbf{x}} \log \pi(\mathbf{x}) \|_2^2 = \mathbb{E}_{\pi} \left[ \frac{1}{2} \| \mathbf{s}(\mathbf{x}, \theta) \|_2^2 + \text{tr}(\nabla_{\mathbf{x}} \mathbf{s}(\mathbf{x}, \theta)) \right] + \text{const}$$

Here  $\nabla_{\mathbf{x}} \mathbf{s}(\mathbf{x}, \theta) = \nabla_{\mathbf{x}}^2 \log p(\mathbf{x}|\theta)$  is a Hessian matrix.

# Score matching

## Theorem

$$\frac{1}{2} \mathbb{E}_\pi \| \mathbf{s}(\mathbf{x}, \boldsymbol{\theta}) - \nabla_{\mathbf{x}} \log \pi(\mathbf{x}) \|_2^2 = \mathbb{E}_\pi \left[ \frac{1}{2} \| \mathbf{s}(\mathbf{x}, \boldsymbol{\theta}) \|_2^2 + \text{tr}(\nabla_{\mathbf{x}} \mathbf{s}(\mathbf{x}, \boldsymbol{\theta})) \right] + \text{const}$$

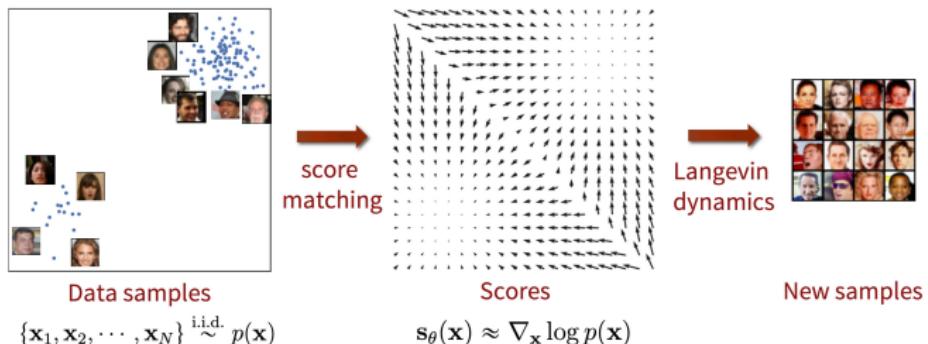
## Proof (only for 1D)

$$\mathbb{E}_\pi \| s(x) - \nabla_x \log \pi(x) \|_2^2 = \mathbb{E}_\pi [s(x)^2 + (\nabla_x \log \pi(x))^2 - 2[s(x) \nabla_x \log \pi(x)]]$$

$$\begin{aligned}\mathbb{E}_\pi [s(x) \nabla_x \log \pi(x)] &= \int \pi(x) \nabla_x \log p(x) \nabla_x \log \pi(x) dx \\ &= \int \nabla_x \log p(x) \nabla_x \pi(x) dx = \pi(x) \nabla_x \log p(x) \Big|_{-\infty}^{+\infty} \\ &\quad - \int \nabla_x^2 \log p(x) \pi(x) dx = -\mathbb{E}_\pi \nabla_x^2 \log p(x)\end{aligned}$$

$$\frac{1}{2} \mathbb{E}_\pi \| s(x) - \nabla_x \log \pi(x) \|_2^2 = \frac{1}{2} \mathbb{E}_\pi [s(x)^2 + \nabla_x s(x)] + \text{const.}$$

# Score matching



Theorem (implicit score matching)

$$\frac{1}{2} \mathbb{E}_\pi \|s(x, \theta) - \nabla_x \log \pi(x)\|_2^2 = \mathbb{E}_\pi \left[ \frac{1}{2} \|s(x, \theta)\|_2^2 + \text{tr}(\nabla_x s(x, \theta)) \right] + \text{const}$$

1. The left hand side is intractable due to unknown  $\pi(x)$  – **denoising score matching**.
2. The right hand side is complex due to Hessian matrix – **sliced score matching**.

## Score matching

Sliced score matching (Hutchinson's trace estimation)

$$\text{tr}(\nabla_{\mathbf{x}} \mathbf{s}(\mathbf{x}, \theta)) = \mathbb{E}_{p(\epsilon)} \left[ \epsilon^T \nabla_{\mathbf{x}} \mathbf{s}(\mathbf{x}, \theta) \epsilon \right],$$

where  $\mathbb{E}[\epsilon] = 0$  and  $\text{Cov}(\epsilon) = \mathbf{I}$ .

Denoising score matching

Let perturb original data by normal noise  $p(\mathbf{x}|\mathbf{x}', \sigma) = \mathcal{N}(\mathbf{x}|\mathbf{x}', \sigma^2 \mathbf{I})$

$$\pi(\mathbf{x}|\sigma) = \int \pi(\mathbf{x}') p(\mathbf{x}|\mathbf{x}', \sigma) d\mathbf{x}'.$$

Then the solution of

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

satisfies  $\mathbf{s}(\mathbf{x}, \theta, \sigma) \approx \mathbf{s}(\mathbf{x}, \theta, 0) = \mathbf{s}(\mathbf{x}, \theta)$  using small enough noise scale  $\sigma$ .

---

*Song Y. Sliced Score Matching: A Scalable Approach to Density and Score Estimation, 2019*

*Vincent P. A connection between score matching and denoising autoencoders. Neural computation, 2011*

# Denoising score matching

## Theorem

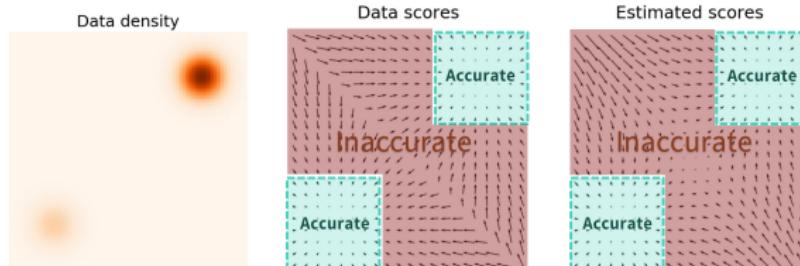
$$\begin{aligned}\mathbb{E}_{\pi(\mathbf{x}|\sigma)} \|\mathbf{s}(\mathbf{x}, \boldsymbol{\theta}, \sigma) - \nabla_{\mathbf{x}} \log \pi(\mathbf{x}|\sigma)\|_2^2 &= \\ &= \mathbb{E}_{\pi(\mathbf{x}')} \mathbb{E}_{p(\mathbf{x}|\mathbf{x}', \sigma)} \|\mathbf{s}(\mathbf{x}, \boldsymbol{\theta}, \sigma) - \nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{x}', \sigma)\|_2^2\end{aligned}$$

Here  $\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{x}', \sigma) = -\frac{\mathbf{x}-\mathbf{x}'}{\sigma^2}$ .

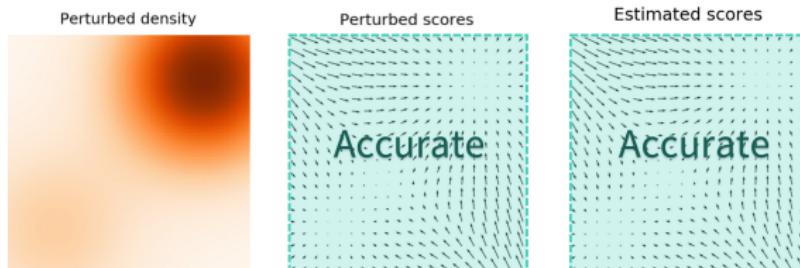
- ▶ The RHS does not need to compute  $\nabla_{\mathbf{x}} \pi(\mathbf{x}|\sigma)$  and even more  $\nabla_{\mathbf{x}} \pi(\mathbf{x})$ .
- ▶  $\mathbf{s}(\mathbf{x}, \boldsymbol{\theta}, \sigma)$  tries to **denoise** a corrupted sample.
- ▶ Score function  $\mathbf{s}(\mathbf{x}, \boldsymbol{\theta}, \sigma)$  parametrized by  $\sigma$ . How to make it?

# Denoising score matching

- If  $\sigma$  is **small**, the score function is not accurate and Langevin dynamics will probably fail to jump between modes.



- If  $\sigma$  is **large**, it is good for low-density regions and multimodal distributions, but we will learn too corrupted distribution.

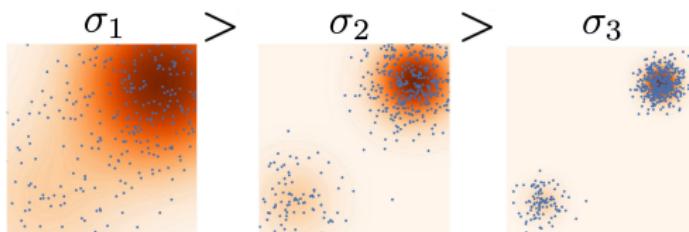


## Noice conditioned score network

- ▶ Define the sequence of noise levels:  $\sigma_1 > \sigma_2 > \dots > \sigma_L$ .
- ▶ Perturb the original data with the different noise level to get  $\pi(\mathbf{x}|\sigma_1), \dots, \pi(\mathbf{x}|\sigma_L)$ .
- ▶ Train denoised score function  $\mathbf{s}(\mathbf{x}, \theta, \sigma)$  for each noise level:

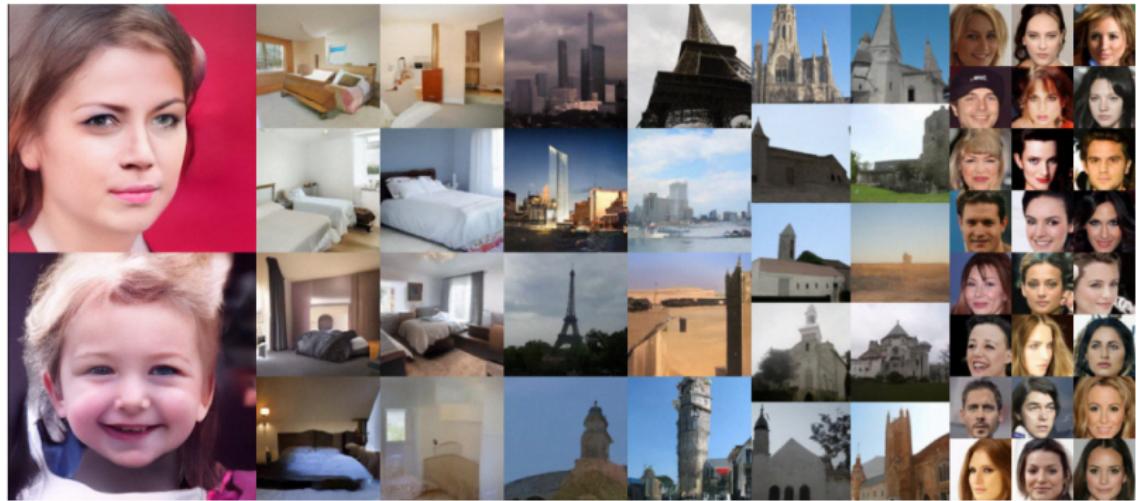
$$\sum_{l=1}^L \sigma_l^2 \mathbb{E}_{\pi(\mathbf{x}')} \mathbb{E}_{p(\mathbf{x}|\mathbf{x}', \sigma_l)} \| \mathbf{s}(\mathbf{x}, \theta, \sigma_l) - \nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{x}', \sigma_l) \|_2^2 \rightarrow \min_{\theta}$$

- ▶ Sample from **annealed** Langevin dynamics (for  $l = 1, \dots, L$ ).



# Noice conditioned score network

## Samples



# Outline

1. Score matching
2. Diffusion models

## Forward diffusion process

Let  $\mathbf{x}_0 \sim \pi(\mathbf{x})$ . Define the Markov chain

$$\mathbf{x}_t = \sqrt{1 - \beta} \mathbf{x}_{t-1} + \sqrt{\beta} \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$$

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t | \sqrt{1 - \beta} \mathbf{x}_{t-1}, \beta \mathbf{I}), \quad \text{where } \beta \in (0, 1)$$

### Statement

Applying the Markov chain to samples from any  $\pi(\mathbf{x})$  we will get  $\mathbf{x}_\infty \sim p_\infty(\mathbf{x}) = \mathcal{N}(0, 1)$ . Here  $p_\infty(\mathbf{x})$  is a **stationary** distribution:

$$p_\infty(\mathbf{x}) = \int q(\mathbf{x} | \mathbf{x}') p_\infty(\mathbf{x}') d\mathbf{x}'.$$

### Statement

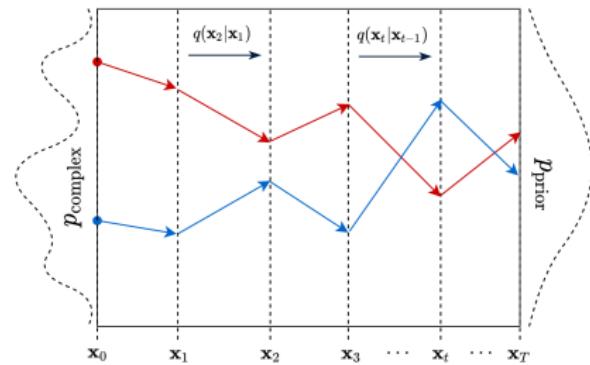
Denote  $\alpha_t = 1 - \beta_t$ ,  $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ . Then

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t | \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I})$$

We could sample from any timestamp using only  $\mathbf{x}_0$ .

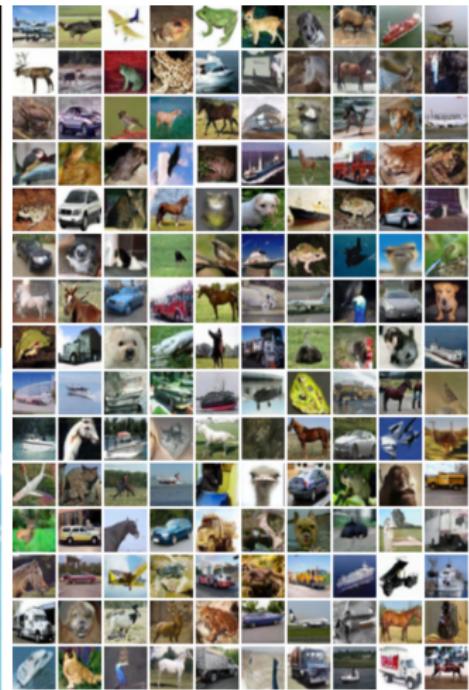
# Diffusion

**Diffusion** refers to the flow of particles from high-density regions towards low-density regions.

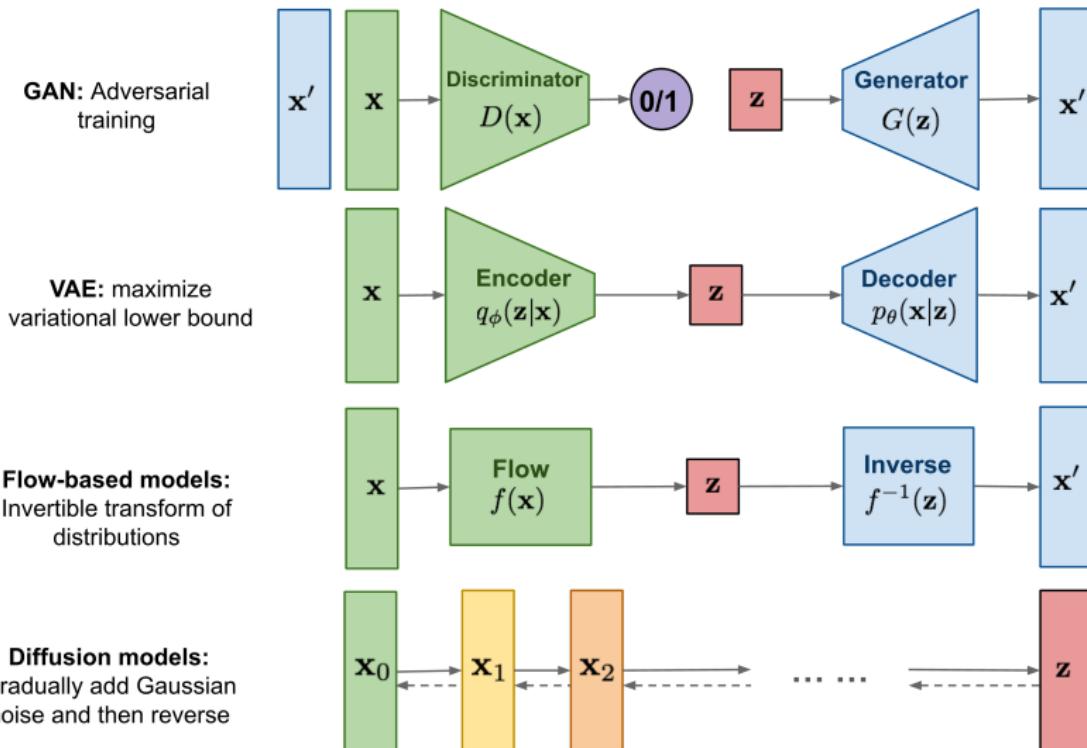


# Denoising diffusion probabilistic model

## Samples



# The poorest course overview :)



## Summary

- ▶ Score matching proposes to minimize Fisher divergence to get score function.
- ▶ Sliced score matching and denoised score matching are two techniques to get scalable algorithm for fitting Fisher divergence.