# Deep Generative Models

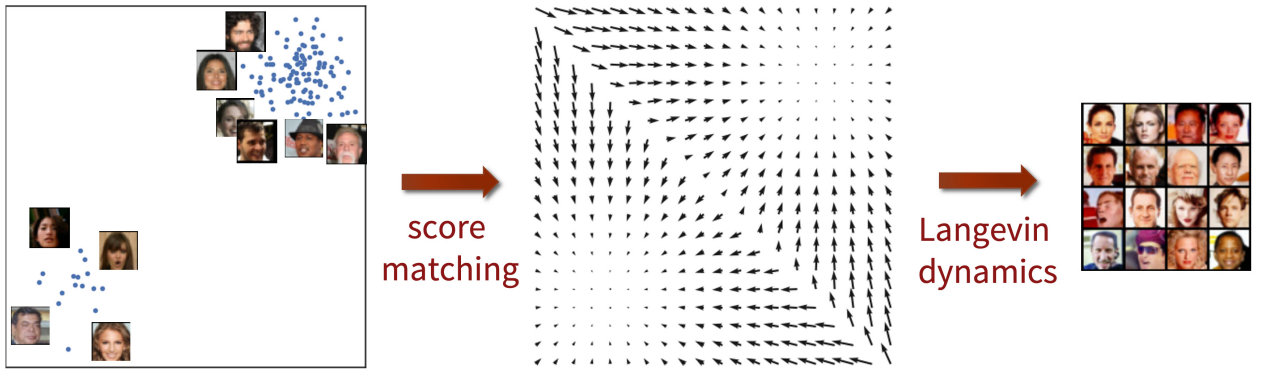
## Lecture 15

Roman Isachenko

Moscow Institute of Physics and Technology

2022 – 2023

# Recap of previous lecture




## Theorem (implicit score matching)

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

1. The left hand side is intractable due to unknown $\pi(\mathbf{x})$ – **denoising score matching**.
2. The right hand side is complex due to Hessian matrix – **sliced score matching (Hutchinson's trace estimation)**.

---

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

# Recap of previous lecture

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

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

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

## Theorem (denoising score matching)

$$\mathbb{E}_{\pi(\mathbf{x}|\sigma)}\left\|\mathbf{s}(\mathbf{x}, \boldsymbol{\theta}, \sigma) - \nabla_{\mathbf{x}} \log \pi(\mathbf{x}|\sigma)\right\|_2^2 =$$

$$= \mathbb{E}_{\pi(\mathbf{x}')}\mathbb{E}_{p(\mathbf{x}|\mathbf{x}', \sigma)}\left\|\mathbf{s}(\mathbf{x}, \boldsymbol{\theta}, \sigma) - \nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{x}', \sigma)\right\|_2^2 + \text{const}(\boldsymbol{\theta})$$

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 $\nabla_{\mathbf{x}} \log \pi(\mathbf{x}|\sigma)$ and even more $\nabla_{\mathbf{x}} \log \pi(\mathbf{x})$.
- $\mathbf{s}(\mathbf{x}, \boldsymbol{\theta}, \sigma)$ tries to **denoise** a corrupted sample.
- Score function $\mathbf{s}(\mathbf{x}, \boldsymbol{\theta}, \sigma)$ parametrized by $\sigma$. How to make it?

---

*Vincent P. A Connection Between Score Matching and Denoising Autoencoders, 2010*

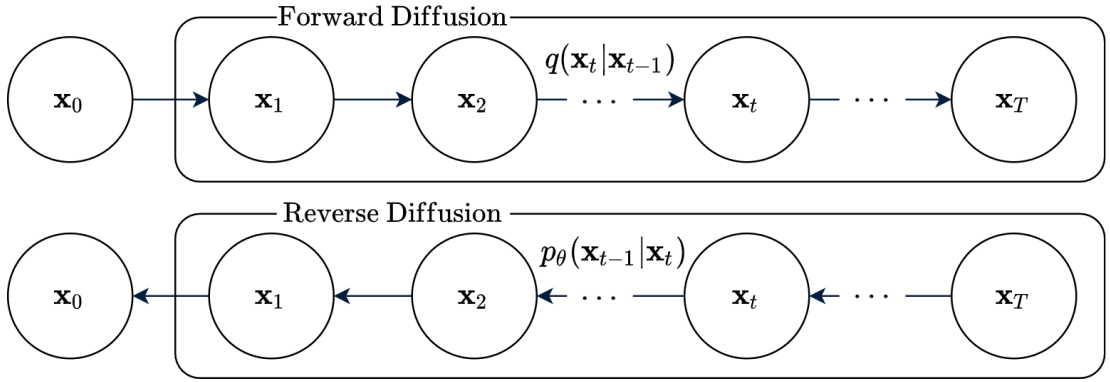
# Outline

1. Denoising diffusion probabilistic model (DDPM)

# Outline

1. Denoising diffusion probabilistic model (DDPM)

# Gaussian diffusion model as VAE




- Let treat $\mathbf{z} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ as a latent variable (**note:** each $\mathbf{x}_t$ has the same size).
- Variational posterior distribution (**note:** there is no learnable parameters)

$$q(\mathbf{z}|\mathbf{x}) = q(\mathbf{x}_1, \ldots, \mathbf{x}_T|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}).$$

- Probabilistic model

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})$$

- Generative distribution and prior

$$p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = p(\mathbf{x}_0|\mathbf{x}_1, \boldsymbol{\theta}); \quad p(\mathbf{z}|\boldsymbol{\theta}) = \prod_{t=2}^{T} p(\mathbf{x}_{t-1}|\mathbf{x}_t, \boldsymbol{\theta})$$

---

*Das A. An introduction to Diffusion Probabilistic Models, blog post, 2021*

# Gaussian diffusion model as VAE

## ELBO

$$\log p(\mathbf{x}|\boldsymbol{\theta}) \geq \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{x})} = \mathcal{L}(q, \boldsymbol{\theta}) \rightarrow \max_{q, \boldsymbol{\theta}}$$

## Statement

$$\mathcal{L}(q, \boldsymbol{\theta}) = \mathbb{E}_{q(\mathbf{x}_1, \ldots, \mathbf{x}_T|\mathbf{x}_0)} \log \frac{p(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_T|\boldsymbol{\theta})}{q(\mathbf{x}_1, \ldots, \mathbf{x}_T|\mathbf{x}_0)} =$$

$$= \mathbb{E}_q \Big[ \log p(\mathbf{x}_0|\mathbf{x}_1, \boldsymbol{\theta}) - \sum_{t=2}^{T} \underbrace{KL\big(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)||p(\mathbf{x}_{t-1}|\mathbf{x}_t, \boldsymbol{\theta})\big)}_{\mathcal{L}_t} -$$

$$- KL\big(q(\mathbf{x}_T|\mathbf{x}_0)||p(\mathbf{x}_T)\big) \Big]$$

- First term is a decoder distribution (could be AR model or discretized distribution (like mixture of logistics)).
- Third term is constant (KL between two standard normals).

*Ho J. Denoising Diffusion Probabilistic Models, 2020*

# Gaussian diffusion model as VAE

$$\mathcal{L}_t = KL\big(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)||p(\mathbf{x}_{t-1}|\mathbf{x}_t, \boldsymbol{\theta})\big).$$

Here

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t\mathbf{I}),$$

$\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0)$ and $\tilde{\beta}_t$ have analytical formulas (we omit it) and both dependent on $\beta_t$.

- **Note:** We do not have an analytical expression for $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$.
- Assume $\boldsymbol{\sigma}^2(\mathbf{x}_t, \boldsymbol{\theta}, t) = \tilde{\beta}_t\mathbf{I}$ (reminder: $p(\mathbf{x}_{t-1}|\mathbf{x}_t, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}_{t-1}|\boldsymbol{\mu}(\mathbf{x}_t, \boldsymbol{\theta}, t), \boldsymbol{\sigma}^2(\mathbf{x}_t, \boldsymbol{\theta}, t))$).
- Use KL formula between two normal distributions:

$$\begin{aligned}\mathcal{L}_t &= KL\Big(\mathcal{N}\big(\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t\mathbf{I}\big)||\mathcal{N}\big(\boldsymbol{\mu}(\mathbf{x}_t, \boldsymbol{\theta}, t), \tilde{\beta}_t\mathbf{I}\big)\Big) \\ &= \mathbb{E}_{\boldsymbol{\epsilon}}\left[\frac{1}{2\tilde{\beta}_t}\big\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}(\mathbf{x}_t, \boldsymbol{\theta}, t)\big\|^2\right]\end{aligned}$$

---

*Ho J. Denoising Diffusion Probabilistic Models, 2020*

# Gaussian diffusion model as VAE

$$\mathcal{L}_t = \mathbb{E}_{\boldsymbol{\epsilon}} \left[ \frac{1}{2\tilde{\beta}_t} \|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}(\mathbf{x}_t, \boldsymbol{\theta}, t)\|^2 \right] =$$

$$= \mathbb{E}_{\boldsymbol{\epsilon}} \left[ \frac{1}{2\tilde{\beta}_t} \left\| \frac{1}{\sqrt{1-\beta_t}} \left( \mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon} \right) - \boldsymbol{\mu}(\mathbf{x}_t, \boldsymbol{\theta}, t) \right\|^2 \right]$$

Here we used the analytic expression for $\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0)$.

## Reparametrization

$$\boldsymbol{\mu}(\mathbf{x}_t, \boldsymbol{\theta}, t) = \frac{1}{\sqrt{1-\beta_t}} \left( \mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}(\mathbf{x}_t, \boldsymbol{\theta}, t) \right)$$

## KL term

$$\mathcal{L}_t = \mathbb{E}_{\boldsymbol{\epsilon}} \left[ \frac{\beta_t^2}{2\tilde{\beta}_t(1-\beta_t)} \left\| \frac{\boldsymbol{\epsilon}}{\sqrt{1-\bar{\alpha}_t}} - \frac{\boldsymbol{\epsilon}(\mathbf{x}_t, \boldsymbol{\theta}, t)}{\sqrt{1-\bar{\alpha}_t}} \right\|^2 \right]$$

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \cdot \mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t} \cdot \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$$

---

*Ho J. Denoising Diffusion Probabilistic Models, 2020*

# Gaussian diffusion model vs Score matching

$$\mathcal{L}_t = \mathbb{E}_{\boldsymbol{\epsilon}} \left[ \frac{\beta_t^2}{2\tilde{\beta}_t(1-\beta_t)} \left\| \frac{\boldsymbol{\epsilon}}{\sqrt{1-\bar{\alpha}_t}} - \frac{\boldsymbol{\epsilon}(\mathbf{x}_t, \boldsymbol{\theta}, t)}{\sqrt{1-\bar{\alpha}_t}} \right\|^2 \right]$$

- Result from Statement 2

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

- Score of noised distribution

$$\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t|\mathbf{x}_0) = -\frac{\boldsymbol{\epsilon}}{\sqrt{1-\bar{\alpha}_t}}, \quad \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}(0,1).$$

- Let reparametrize our model:

$$\mathbf{s}(\mathbf{x}_t, \boldsymbol{\theta}, t) = \frac{\boldsymbol{\epsilon}(\mathbf{x}_t, \boldsymbol{\theta}, t)}{\sqrt{1-\bar{\alpha}_t}}.$$

## Noise conditioned score network

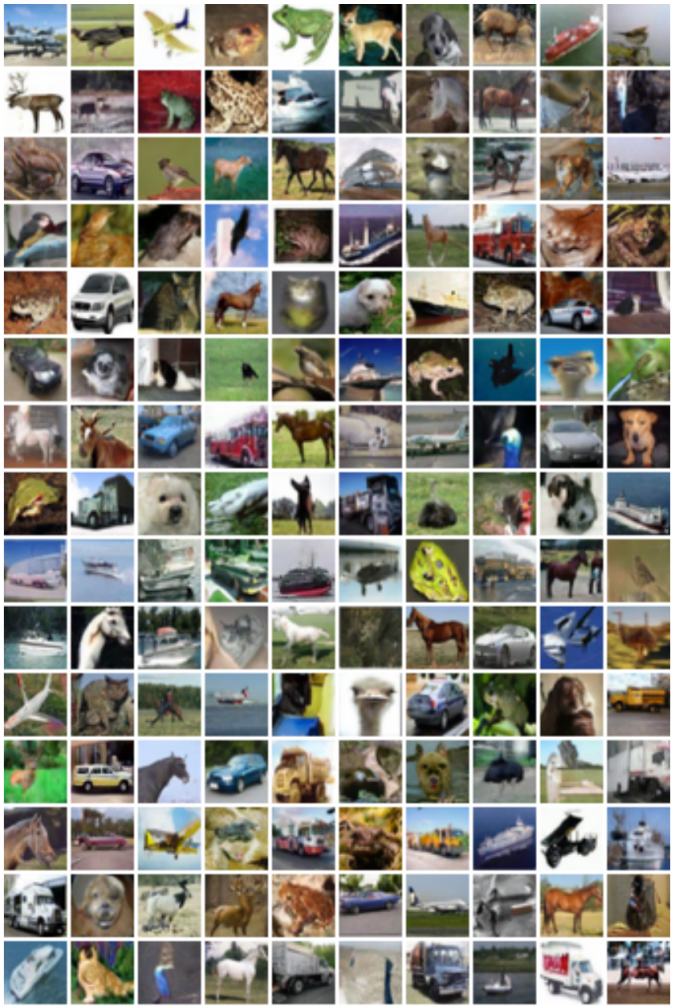
$$\mathbb{E}_{p(\mathbf{x}|\mathbf{x}',\sigma_l)} \left\| \mathbf{s}(\mathbf{x}, \boldsymbol{\theta}, \sigma_l) - \nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{x}', \sigma_l) \right\|_2^2 \to \min_{\boldsymbol{\theta}}$$

---

*Ho J. Denoising Diffusion Probabilistic Models, 2020*

# Denoising diffusion probabilistic model (DDPM)

## Samples



*Ho J. Denoising Diffusion Probabilistic Models, 2020*

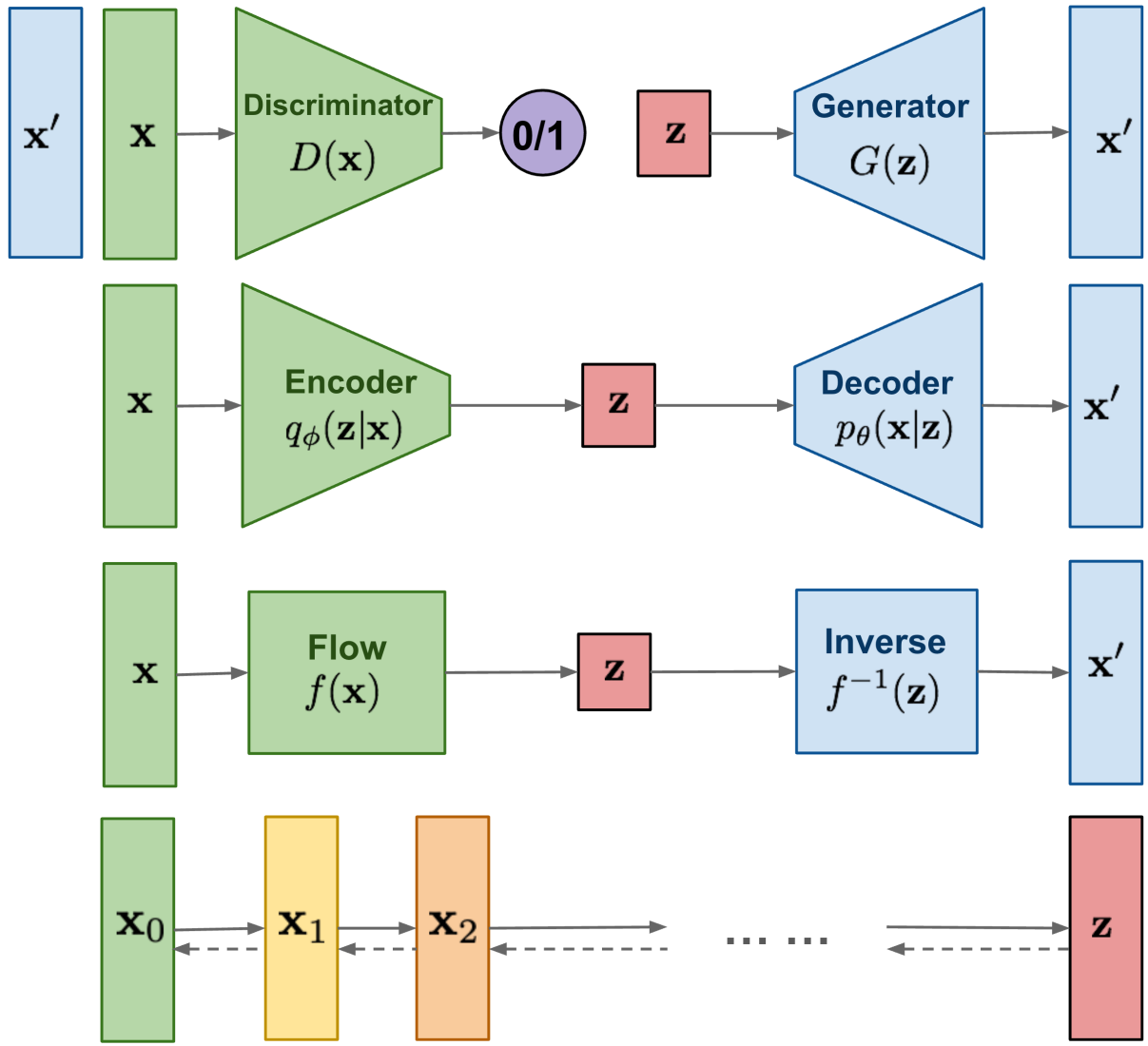
# The poorest course overview :)

**GAN:** Adversarial training

**VAE:** maximize variational lower bound

**Flow-based models:** Invertible transform of distributions

**Diffusion models:** Gradually add Gaussian noise and then reverse




*Weng L. What are Diffusion Models?, blog post, 2021*

# Summary

- Diffusion model is a VAE model which reverts gaussian diffusion process using variational inference.

- Objective of diffusion model is closely related to the noise conditioned score network and score matching.