

# Deep Generative Models

## Lecture 12

Roman Isachenko



2024, Spring

## Recap of previous lecture

### Forward gaussian diffusion process

Let  $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x})$ ,  $\beta_t \in (0, 1)$ ,  $\alpha_t = 1 - \beta_t$  and  $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ .

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

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

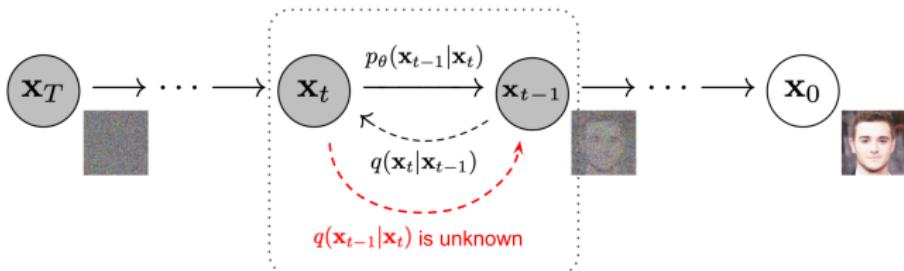
$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t | \sqrt{1 - \beta_t} \cdot \mathbf{x}_{t-1}, \beta_t \cdot \mathbf{I});$$

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

1.  $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x})$ ;
2.  $\mathbf{x}_t = \sqrt{1 - \beta_t} \cdot \mathbf{x}_{t-1} + \sqrt{\beta_t} \cdot \boldsymbol{\epsilon}$ , where  $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$ ,  $t \geq 1$ ;
3.  $\mathbf{x}_T \sim p_\infty(\mathbf{x}) = \mathcal{N}(0, 1)$ , where  $T \gg 1$ .

If we are able to invert this process, we will get the way to sample  $\mathbf{x} \sim \pi(\mathbf{x})$  using noise samples  $p_\infty(\mathbf{x}) = \mathcal{N}(0, \mathbf{I})$ .

## Recap of previous lecture



### Reverse gaussian diffusion process

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t) = \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})q(\mathbf{x}_{t-1})}{q(\mathbf{x}_t)}$$

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

- ▶  $q(\mathbf{x}_{t-1})$ ,  $q(\mathbf{x}_t)$  are intractable.
- ▶ If  $\beta_t$  is small enough,  $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$  will be Gaussian (Feller, 1949).
- ▶  $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$  defines how to denoise a noisy image  $\mathbf{x}_t$  with access to the completely denoised image  $\mathbf{x}_0$ .

# Recap of previous lecture

Let's define the reverse process

$$q(\mathbf{x}_{t-1} | \mathbf{x}_t) \approx p(\mathbf{x}_{t-1} | \mathbf{x}_t, \theta) = \mathcal{N}(\mathbf{x}_{t-1} | \mu_\theta(\mathbf{x}_t, t), \sigma_\theta^2(\mathbf{x}_t, t))$$

Forward process

1.  $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x})$ ;

2.  $\mathbf{x}_t = \sqrt{1 - \beta_t} \cdot \mathbf{x}_{t-1} + \sqrt{\beta_t} \cdot \epsilon$ ,  
where  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ ,  $t \geq 1$ ;

3.  $\mathbf{x}_T \sim p_\infty(\mathbf{x}) = \mathcal{N}(0, \mathbf{I})$ .

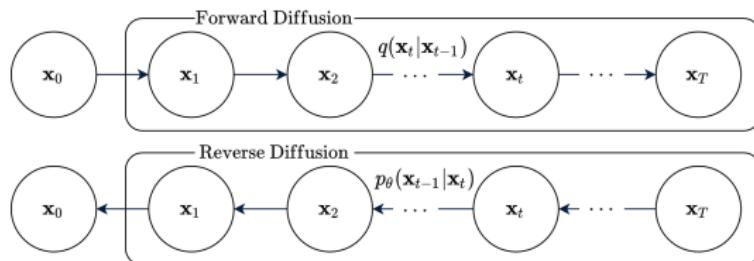
Reverse process

1.  $\mathbf{x}_T \sim p_\infty(\mathbf{x}) = \mathcal{N}(0, \mathbf{I})$ ;

2.  $\mathbf{x}_{t-1} = \sigma_\theta(\mathbf{x}_t, t) \cdot \epsilon + \mu_\theta(\mathbf{x}_t, t)$ ;

3.  $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x})$ ;

Gaussian diffusion model as VAE



## Recap of previous lecture

- ▶  $\mathbf{z} = (\mathbf{x}_1, \dots, \mathbf{x}_T)$  is a latent variable.
- ▶ Variational posterior distribution

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

- ▶ Generative distribution and prior

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

## ELBO

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

$$\begin{aligned} \mathcal{L}(q, \boldsymbol{\theta}) &= \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)} \log p(\mathbf{x}_0|\mathbf{x}_1, \boldsymbol{\theta}) - \textcolor{violet}{KL}(q(\mathbf{x}_T|\mathbf{x}_0) || p(\mathbf{x}_T)) - \\ &\quad - \sum_{t=2}^T \underbrace{\mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} \textcolor{violet}{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) || p(\mathbf{x}_{t-1}|\mathbf{x}_t, \boldsymbol{\theta}))}_{\mathcal{L}_t} \end{aligned}$$

## Recap of previous lecture

ELBO of gaussian diffusion model

$$\begin{aligned}\mathcal{L}(q, \theta) = & \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)} \log p(\mathbf{x}_0|\mathbf{x}_1, \theta) - KL(q(\mathbf{x}_T|\mathbf{x}_0)||p(\mathbf{x}_T)) - \\ & - \sum_{t=2}^T \underbrace{\mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} KL(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)||p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta))}_{\mathcal{L}_t}\end{aligned}$$

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

$$p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta) = \mathcal{N}(\mathbf{x}_{t-1}|\boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \sigma_\theta^2(\mathbf{x}_t, t))$$

Our assumption:  $\sigma_\theta^2(\mathbf{x}_t, t) = \tilde{\beta}_t \mathbf{I}$ .

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

# Outline

## 1. Denoising Diffusion Probabilistic Model (DDPM)

Reparametrization of gaussian diffusion model

Overview of DDPM

## 2. DDPM vs NCSN

# Outline

## 1. Denoising Diffusion Probabilistic Model (DDPM)

Reparametrization of gaussian diffusion model

Overview of DDPM

## 2. DDPM vs NCSN

# Outline

## 1. Denoising Diffusion Probabilistic Model (DDPM)

Reparametrization of gaussian diffusion model

Overview of DDPM

## 2. DDPM vs NCSN

# Reparametrization of DDPM

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

$\mathcal{L}_t$  is the mean of KL between two normal distributions:

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

$$p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta) = \mathcal{N}(\mathbf{x}_{t-1} | \mu_\theta(\mathbf{x}_t, t), \sigma_\theta^2(\mathbf{x}_t, t) \mathbf{I})$$

Here

$$\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)}{1 - \bar{\alpha}_t} \mathbf{x}_0;$$

$$\tilde{\beta}_t = \frac{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} = \text{const.}$$

Let assume

$$\sigma_\theta^2(\mathbf{x}_t, t) = \tilde{\beta}_t \mathbf{I} \quad \Rightarrow \quad p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta) = \mathcal{N}(\mathbf{x}_{t-1} | \mu_\theta(\mathbf{x}_t, t), \tilde{\beta}_t \mathbf{I}).$$

# Reparametrization of DDPM

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

Use the formula for KL between two normal distributions:

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

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon} \quad \Rightarrow \quad \mathbf{x}_0 = \frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}}{\sqrt{\bar{\alpha}_t}}$$

$$\begin{aligned}\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) &= \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)}{1 - \bar{\alpha}_t} \mathbf{x}_0 \\ &= \frac{1}{\sqrt{\alpha_t}} \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{\alpha_t(1 - \bar{\alpha}_t)}} \boldsymbol{\epsilon}\end{aligned}$$

# Reparametrization of DDPM

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

## Reparametrization

$$\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) = \frac{1}{\sqrt{\alpha_t}} \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{\alpha_t(1 - \bar{\alpha}_t)}} \epsilon$$

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{\alpha_t(1 - \bar{\alpha}_t)}} \epsilon_\theta(\mathbf{x}_t, t)$$

$$\begin{aligned} \mathcal{L}_t &= \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \mathbf{I})} \left[ \frac{(1 - \alpha_t)^2}{2\tilde{\beta}_t \alpha_t (1 - \bar{\alpha}_t)} \|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2 \right] \\ &= \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \mathbf{I})} \left[ \frac{(1 - \alpha_t)^2}{2\tilde{\beta}_t \alpha_t (1 - \bar{\alpha}_t)} \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t)\|^2 \right] \end{aligned}$$

At each step of reverse diffusion process we try to predict the noise  $\epsilon$  that we used in the forward diffusion process!

# Reparametrization of DDPM

$$\begin{aligned}\mathcal{L}(q, \theta) = & \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)} \log p(\mathbf{x}_0|\mathbf{x}_1, \theta) - KL(q(\mathbf{x}_T|\mathbf{x}_0)||p(\mathbf{x}_T)) - \\ & - \sum_{t=2}^T \underbrace{\mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} KL(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)||p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta))}_{\mathcal{L}_t}\end{aligned}$$

$$\mathcal{L}_t = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \mathbf{I})} \left[ \frac{(1 - \alpha_t)^2}{2\tilde{\beta}_t \alpha_t (1 - \bar{\alpha}_t)} \left\| \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t) \right\|^2 \right]$$

Simplified objective

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t \sim U[2, T]} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \mathbf{I})} \left\| \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t) \right\|^2$$

# Outline

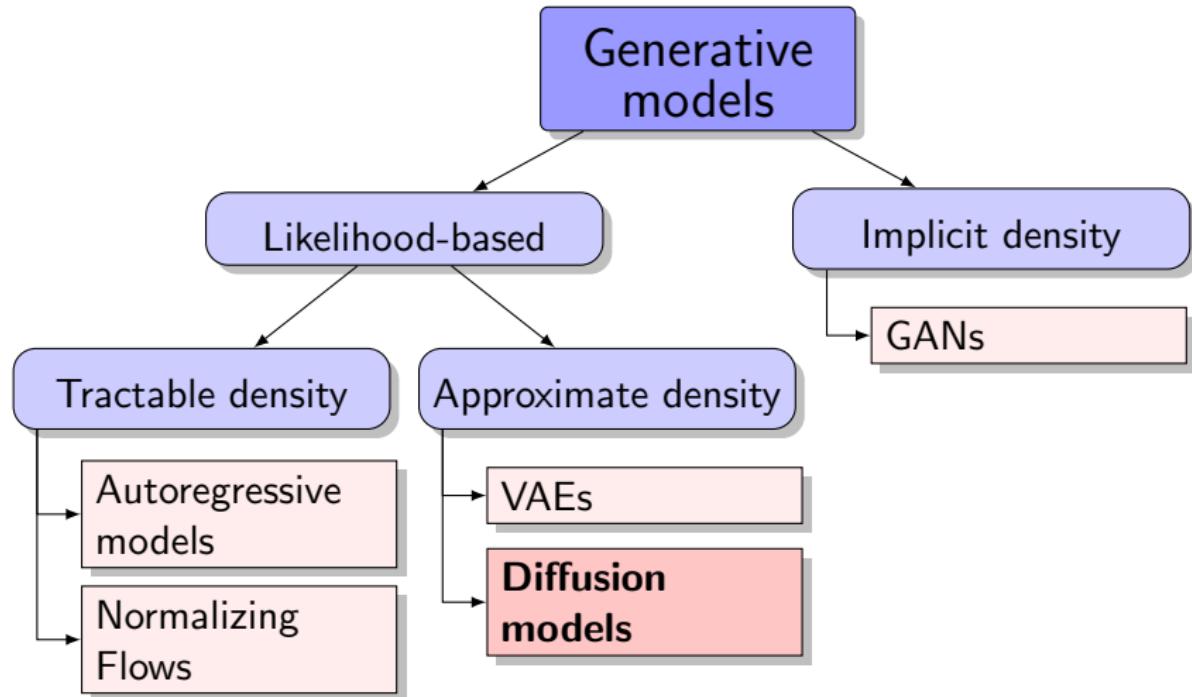
## 1. Denoising Diffusion Probabilistic Model (DDPM)

Reparametrization of gaussian diffusion model

Overview of DDPM

## 2. DDPM vs NCSN

# Generative models zoo



# Denoising diffusion probabilistic model (DDPM)

DDPM is a VAE model

- ▶ Encoder is a fixed Gaussian Markov chain  $q(\mathbf{x}_1, \dots, \mathbf{x}_T | \mathbf{x}_0)$ .
- ▶ Latent variable is a hierarchical (in each step the dim. of the latent equals to the dim of the input).
- ▶ Decoder is a simple Gaussian model  $p(\mathbf{x}_0 | \mathbf{x}_1, \theta)$ .
- ▶ Prior distribution is given by parametric Gaussian Makov chain  $p(\mathbf{x}_{t-1} | \mathbf{x}_t, \theta)$ .

Forward process

1.  $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x})$ ;
2.  $\mathbf{x}_t = \sqrt{1 - \beta_t} \cdot \mathbf{x}_{t-1} + \sqrt{\beta_t} \cdot \boldsymbol{\epsilon}$ ,  
where  $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ ,  $t \geq 1$ ;
3.  $\mathbf{x}_T \sim p_\infty(\mathbf{x}) = \mathcal{N}(0, \mathbf{I})$ .

Reverse process

1.  $\mathbf{x}_T \sim p_\infty(\mathbf{x}) = \mathcal{N}(0, \mathbf{I})$ ;
2.  $\mathbf{x}_{t-1} = \sigma_\theta(\mathbf{x}_t, t) \cdot \boldsymbol{\epsilon} + \mu_\theta(\mathbf{x}_t, t)$ ;
3.  $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x})$ ;

# Denoising diffusion probabilistic model (DDPM)

## Training

1. Get the sample  $\mathbf{x}_0 \sim \pi(\mathbf{x})$ .
2. Sample timestamp  $t \sim U[1, T]$  and the noise  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ .
3. Get noisy image  $\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$ .
4. Compute loss  $\mathcal{L}_{\text{simple}} = \|\epsilon - \epsilon_{\theta}(\mathbf{x}_t, t)\|^2$ .

## Sampling

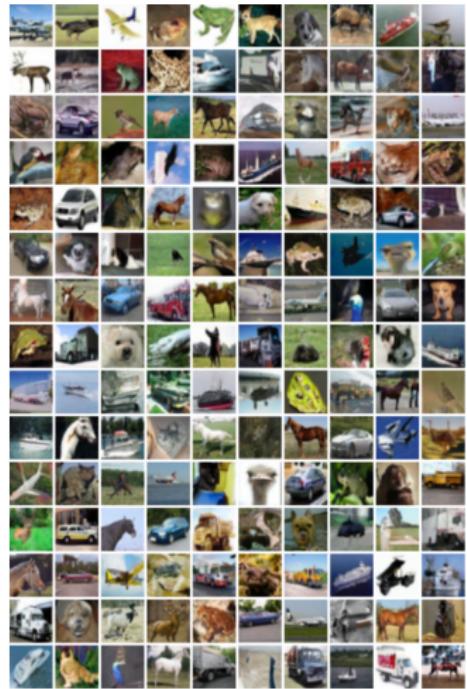
1. Sample  $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ .
2. Compute mean of  $p(\mathbf{x}_{t-1} | \mathbf{x}_t, \theta) = \mathcal{N}(\mu_{\theta}(\mathbf{x}_t, t), \tilde{\beta}_t \mathbf{I})$ :

$$\mu_{\theta}(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{\alpha_t(1 - \bar{\alpha}_t)}} \epsilon_{\theta}(\mathbf{x}_t, t)$$

3. Get denoised image  $\mathbf{x}_{t-1} = \mu_{\theta}(\mathbf{x}_t, t) + \sqrt{\tilde{\beta}_t} \cdot \epsilon$ , where  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ .

# Denoising diffusion probabilistic model (DDPM)

## Samples



# Outline

## 1. Denoising Diffusion Probabilistic Model (DDPM)

Reparametrization of gaussian diffusion model

Overview of DDPM

## 2. DDPM vs NCSN

# DDPM vs NCSN

NCSN objective

$$\mathbb{E}_{p(\mathbf{x}'|\mathbf{x}, \sigma_I)} \| \mathbf{s}_\theta(\mathbf{x}', \sigma_I) - \nabla_{\mathbf{x}'} \log p(\mathbf{x}'|\mathbf{x}, \sigma_I) \|_2^2 \rightarrow \min_\theta$$

DDPM objective

$$\mathcal{L}_t = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \mathbf{I})} \left[ \frac{(1 - \alpha_t)^2}{2\tilde{\beta}_t \alpha_t} \left\| \frac{\epsilon}{\sqrt{1 - \bar{\alpha}_t}} - \frac{\epsilon_\theta(\mathbf{x}_t, t)}{\sqrt{1 - \bar{\alpha}_t}} \right\|^2 \right]$$

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

$$\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t | \mathbf{x}_0) = -\frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t} \cdot \mathbf{x}_0}{1 - \bar{\alpha}_t} = -\frac{\epsilon}{\sqrt{1 - \bar{\alpha}_t}}.$$

Let reparametrize our model:

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

## Summary

- ▶ DDPM is a VAE model with hierarchical latent variables.
- ▶ At each step DDPM predicts the noise that was used in the forward diffusion process.
- ▶ DDPM is really slow, because we have to apply the model  $T$  times.
- ▶ Objective of DDPM is closely related to the noise conditioned score network and score matching.