

Deep Generative Models

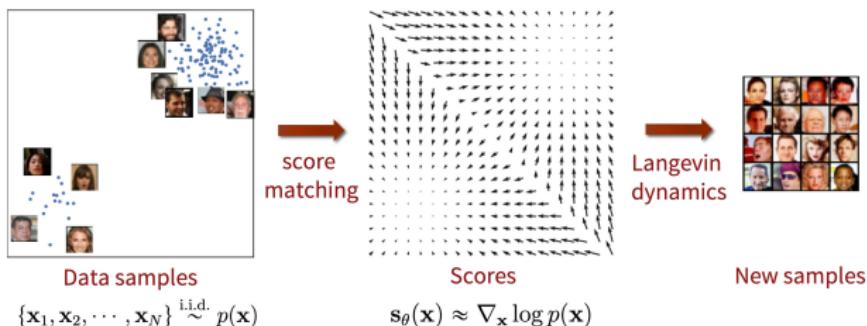
Lecture 15

Roman Isachenko

Moscow Institute of Physics and Technology

2022 – 2023

Recap of previous lecture



Theorem (implicit score matching)

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

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)**.

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

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

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

Theorem (denoising score matching)

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

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

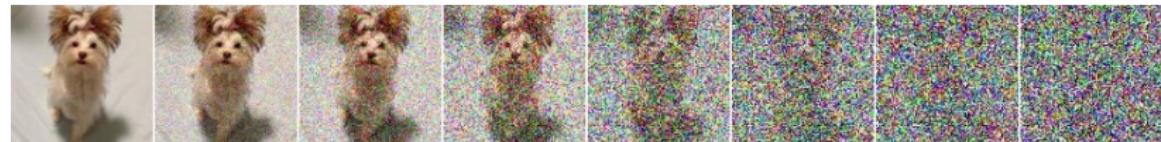
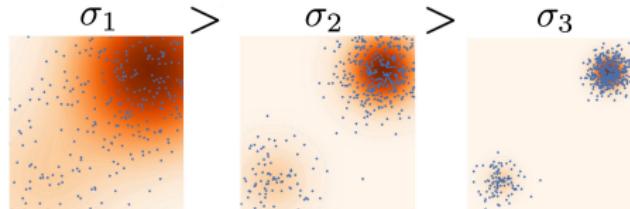
Recap of previous lecture

Noise conditioned score network

- ▶ Define the sequence of noise levels: $\sigma_1 > \sigma_2 > \dots > \sigma_L$.
- ▶ Train denoised score function $s(\mathbf{x}', \theta, \sigma)$ for each noise level:

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

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



Recap of previous lecture

Forward gaussian diffusion process

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

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

- ▶ $p_\infty(\mathbf{x}) = \mathcal{N}(0, 1)$
- ▶ $q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t | \sqrt{\bar{\alpha}_t} \cdot \mathbf{x}_0, (1 - \bar{\alpha}_t) \cdot \mathbf{I})$

Reverse gaussian diffusion process

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

Forward process

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

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

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

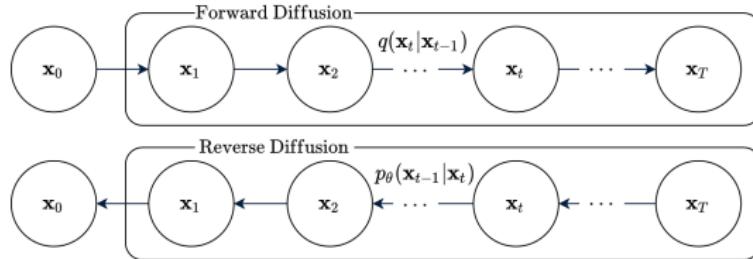
Reverse process

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

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

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

Recap of previous lecture



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

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

- ▶ 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

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

Outline

1. Denoising diffusion probabilistic model (DDPM)
 - Objective of DDPM
 - Reparametrization of DDPM
 - Overview
2. The worst course overview

Outline

1. Denoising diffusion probabilistic model (DDPM)

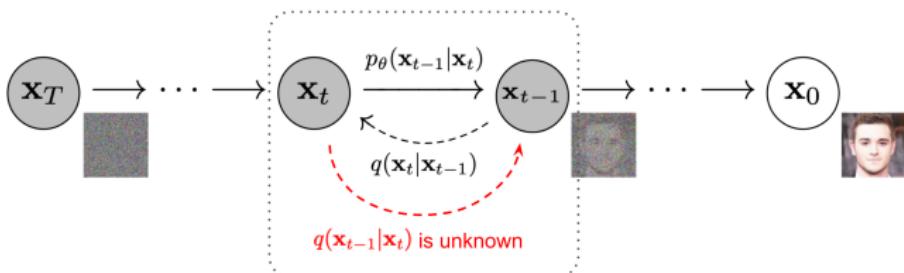
Objective of DDPM

Reparametrization of DDPM

Overview

2. The worst course overview

Reverse gaussian diffusion process



Forward process

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

Reverse process

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

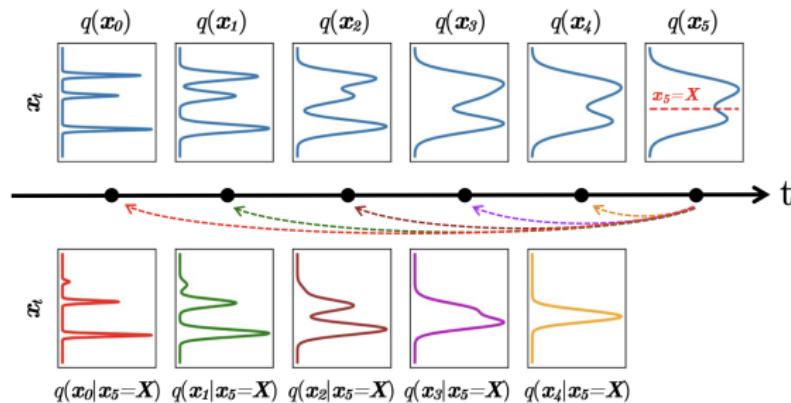
- ▶ $q(\mathbf{x}_{t-1})$, $q(\mathbf{x}_t)$ are intractable.
- ▶ If β_t is small enough, $q(\mathbf{x}_{t-1} | \mathbf{x}_t)$ will be Gaussian (Feller, 1949).

Feller W. On the theory of stochastic processes, with particular reference to applications, 1949

Reverse gaussian diffusion process

Reverse process

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



Important distribution

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

Outline

1. Denoising diffusion probabilistic model (DDPM)

Objective of DDPM

Reparametrization of DDPM

Overview

2. The worst course overview

Objective of DDPM

ELBO

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

Derivation

$$\begin{aligned}\mathcal{L}(q, \theta) &= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \log \frac{p(\mathbf{x}_0, \mathbf{x}_{1:T}|\theta)}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} = \mathbb{E}_q \log \frac{\prod_{t=1}^T p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta) p(\mathbf{x}_T)}{\prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1})} \\ &= \mathbb{E}_q \left[\log p(\mathbf{x}_T) + \log \frac{p(\mathbf{x}_0|\mathbf{x}_1, \theta)}{q(\mathbf{x}_1|\mathbf{x}_0)} + \sum_{t=2}^T \log \left(\frac{p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right) \right] \\ q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) &= \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1}) q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \\ \log \left(\frac{p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right) &= \log \left(\frac{p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta) q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) q(\mathbf{x}_t|\mathbf{x}_0)} \right)\end{aligned}$$

Objective of DDPM

Derivation

$$\begin{aligned}\mathcal{L}(q, \theta) &= \mathbb{E}_q \left[\log p(\mathbf{x}_T) + \log \frac{p(\mathbf{x}_0|\mathbf{x}_1, \theta)}{q(\mathbf{x}_1|\mathbf{x}_0)} + \sum_{t=2}^T \log \left(\frac{p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right) \right] \\ &= \mathbb{E}_q \left[\log p(\mathbf{x}_T) + \log \frac{p(\mathbf{x}_0|\mathbf{x}_1, \theta)}{q(\mathbf{x}_1|\mathbf{x}_0)} + \sum_{t=2}^T \log \left(\frac{p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} \right) + \right. \\ &\quad \left. + \sum_{t=2}^T \log \left(\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \right) \right] = \mathbb{E}_q \left[\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)} + \log \frac{p(\mathbf{x}_0|\mathbf{x}_1, \theta)}{q(\mathbf{x}_1|\mathbf{x}_0)} + \right. \\ &\quad \left. + \sum_{t=2}^T \log \left(\frac{p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} \right) \right] = \mathbb{E}_q \left[-KL(q(\mathbf{x}_T|\mathbf{x}_0)||p(\mathbf{x}_T)) + \right. \\ &\quad \left. + \log p(\mathbf{x}_0|\mathbf{x}_1, \theta) - \sum_{t=2}^T KL(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)||p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta)) \right]\end{aligned}$$

Objective of DDPM

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

- ▶ First term is a decoder distribution

$$\log p(\mathbf{x}_0 | \mathbf{x}_1, \theta) = \log \mathcal{N}(\mathbf{x}_0 | \mu(\mathbf{x}_1, \theta, t), \sigma^2(\mathbf{x}_1, \theta, t))$$

- ▶ Second term is constant ($p(\mathbf{x}_T)$ is a standard Normal, $q(\mathbf{x}_T | \mathbf{x}_0)$ is a non-parametrical Normal).
- ▶ \mathcal{L}_t is a KL between two normal distributions:

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

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

Outline

1. Denoising diffusion probabilistic model (DDPM)

Objective of DDPM

Reparametrization of DDPM

Overview

2. The worst course overview

Gaussian diffusion model as VAE

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

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

- ▶ Assume $\sigma^2(\mathbf{x}_t, \theta, t) = \tilde{\beta}_t \mathbf{I}$.
- ▶ Use KL formula between two normal distributions:

$$\mathcal{L}_t = KL\left(\mathcal{N}(\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I}) || \mathcal{N}(\mu(\mathbf{x}_t, \theta, t), \tilde{\beta}_t \mathbf{I})\right)$$

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

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

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

Reparametrization

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

Reparametrization of DDPM

KL term

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

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

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

Gaussian diffusion model vs Score matching

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

- ▶ 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

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

- ▶ Let reparametrize our model:

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

Noise conditioned score network

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

Outline

1. Denoising diffusion probabilistic model (DDPM)

Objective of DDPM

Reparametrization of DDPM

Overview

2. The worst course overview

Denoising diffusion probabilistic model (DDPM)

DDPM is a VAE model

- ▶ Encoder is a fixed Gaussian Markov chain.
- ▶ 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.
- ▶ Prior distribution is given by parametric Gaussian Makov chain.

Forward process

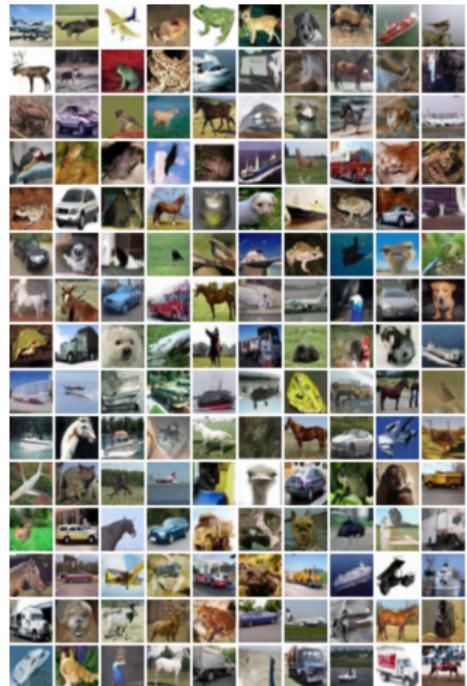
1. $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x});$
2. $\mathbf{x}_t = \sqrt{1 - \beta} \cdot \mathbf{x}_{t-1} + \sqrt{\beta} \cdot \boldsymbol{\epsilon},$
where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1), t \geq 1;$
3. $\mathbf{x}_T \sim p_\infty(\mathbf{x}) = \mathcal{N}(0, 1).$

Reverse process

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

Denoising diffusion probabilistic model (DDPM)

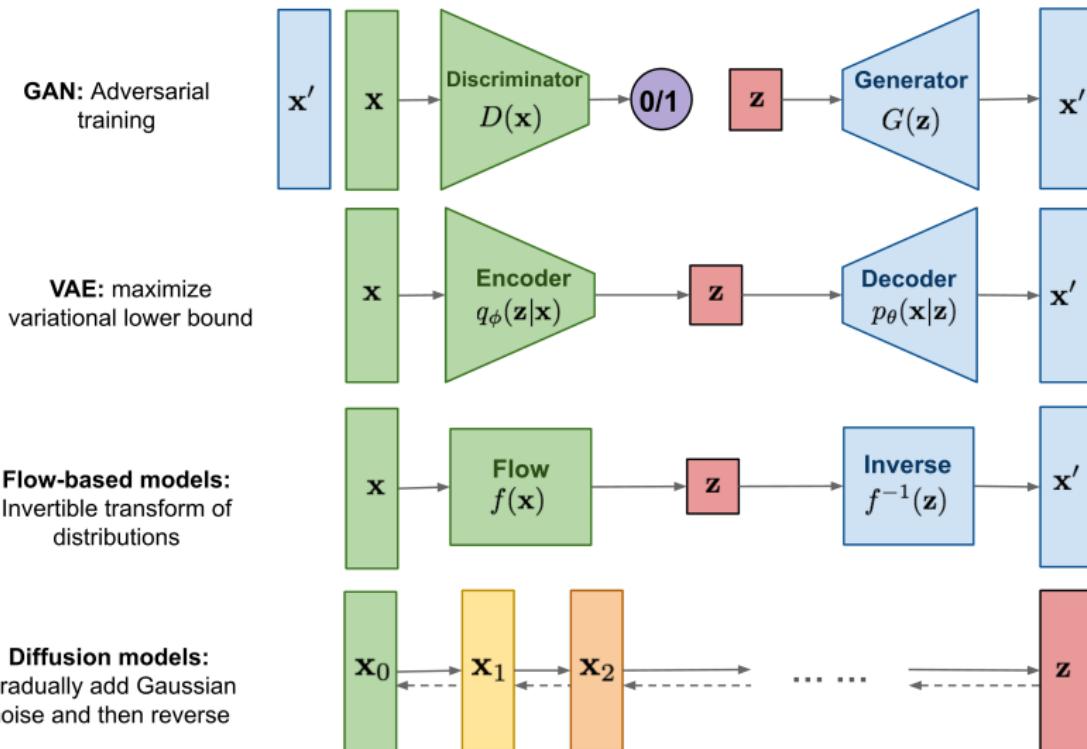
Samples



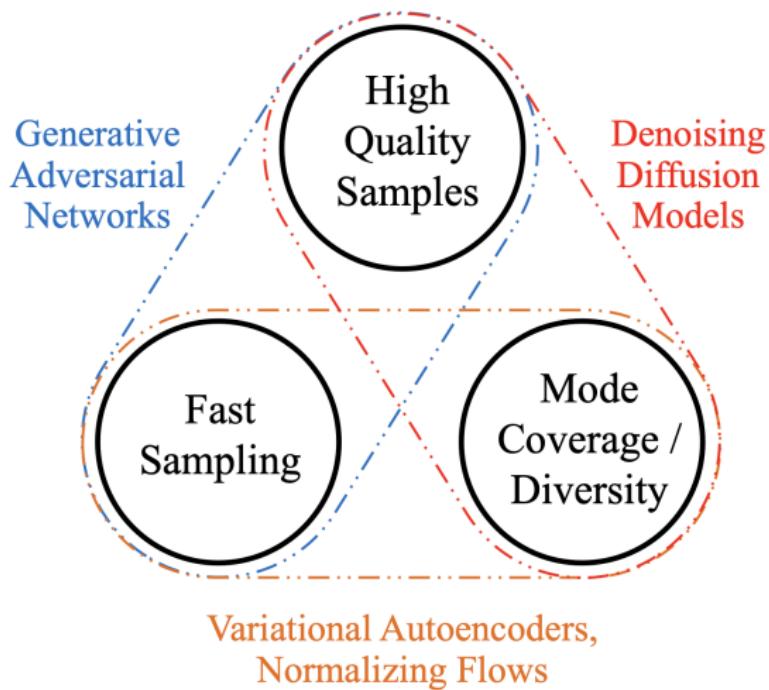
Outline

1. Denoising diffusion probabilistic model (DDPM)
 - Objective of DDPM
 - Reparametrization of DDPM
 - Overview
2. The worst course overview

The worst course overview :)



The worst course overview :)



Summary

- ▶ Diffusion model is a VAE model which reverts gaussian diffusion process using variational inference.
- ▶ ELBO of DDPM is a sum of KL terms.
- ▶ At each step DDPM predicts the noise used in forward process.
- ▶ Objective of DDPM is closely related to the noise conditioned score network and score matching.