

Deep Generative Models

Lecture 14

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)

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

Langevin dynamics

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

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

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. Noise conditioned score network
3. Diffusion models

Outline

1. Score matching
2. Noise conditioned score network
3. 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

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

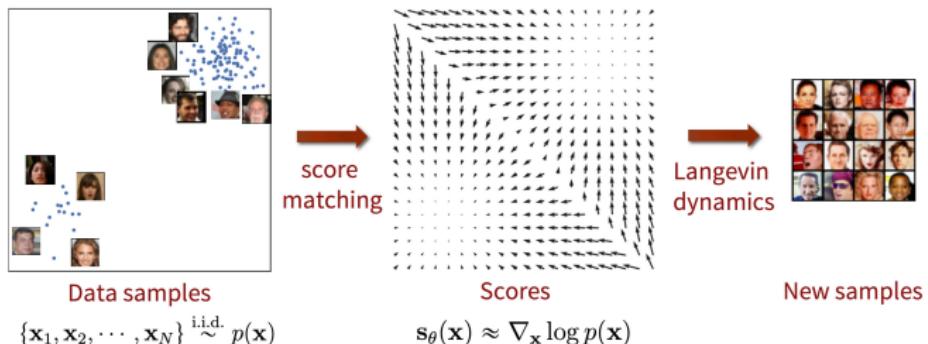
Proof (only for 1D)

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

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

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

Score matching



Theorem (implicit score matching)

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

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**.

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 σ .

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

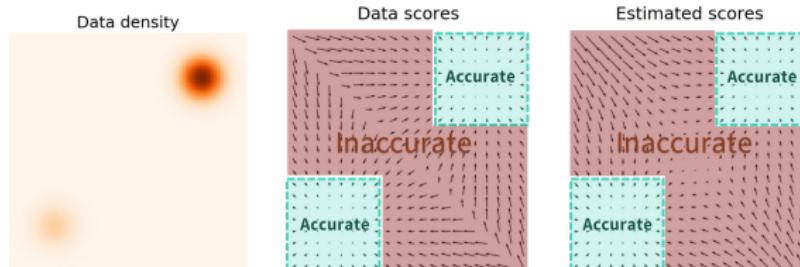
$$\begin{aligned}\mathbb{E}_{\pi(\mathbf{x}|\sigma)} \left\| \mathbf{s}(\mathbf{x}, \boldsymbol{\theta}, \sigma) - \nabla_{\mathbf{x}} \log \pi(\mathbf{x}|\sigma) \right\|_2^2 &= \\ &= \mathbb{E}_{\pi(\mathbf{x}')} \mathbb{E}_{p(\mathbf{x}|\mathbf{x}', \sigma)} \left\| \mathbf{s}(\mathbf{x}, \boldsymbol{\theta},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

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

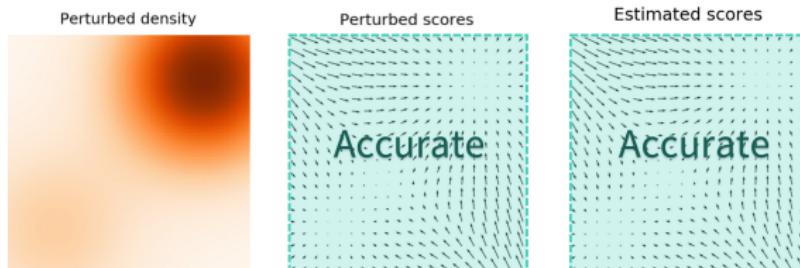
- ▶ The RHS does not need to compute $\nabla_{\mathbf{x}} \log \pi(\mathbf{x}|\sigma)$ and even more $\nabla_{\mathbf{x}} \log \pi(\mathbf{x})$.
- ▶ $\mathbf{s}(\mathbf{x}, \boldsymbol{\theta}, \sigma)$ tries to **denoise** a corrupted sample.
- ▶ Score function $\mathbf{s}(\mathbf{x}, \boldsymbol{\theta},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

Denoising score matching

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



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



Outline

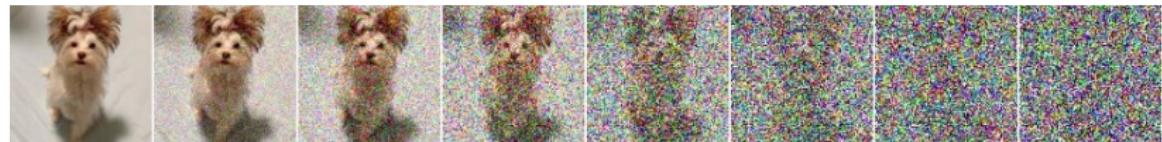
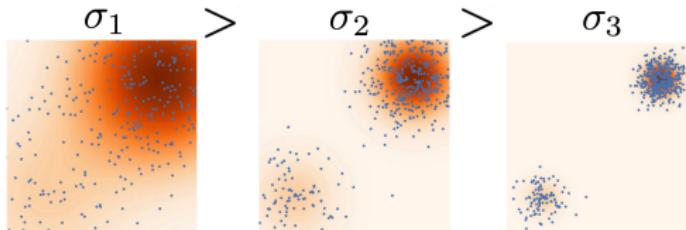
1. Score matching
2. Noise conditioned score network
3. Diffusion models

Noise 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:

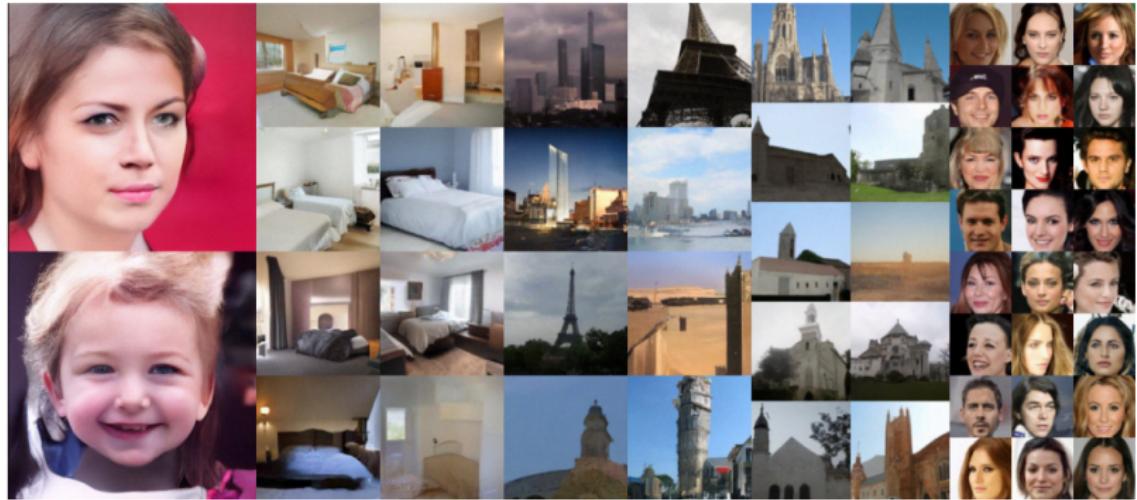
$$\sum_{l=1}^L \sigma_l^2 \mathbb{E}_{\pi(\mathbf{x}')} \mathbb{E}_{p(\mathbf{x}|\mathbf{x}', \sigma_l)} \| \mathbf{s}(\mathbf{x}, \theta, \sigma_l) - \nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{x}', \sigma_l) \|_2^2 \rightarrow \min_{\theta}$$

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



Noise conditioned score network

Samples



Outline

1. Score matching
2. Noise conditioned score network
3. Diffusion models

Forward diffusion process

Let $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x})$, $\beta \in (0, 1)$. Define the Markov chain

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

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

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

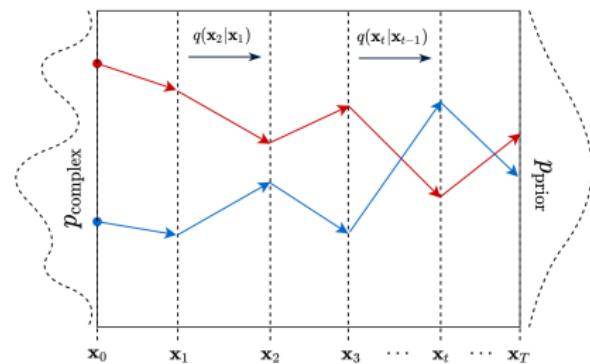
$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$$

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

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

Forward diffusion process

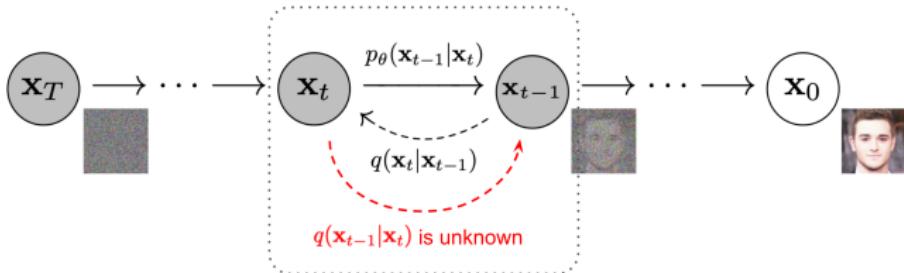
Diffusion refers to the flow of particles from high-density regions towards low-density regions.



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

Now our goal is to revert this process.

Reverse diffusion process



Let define the reverse process

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

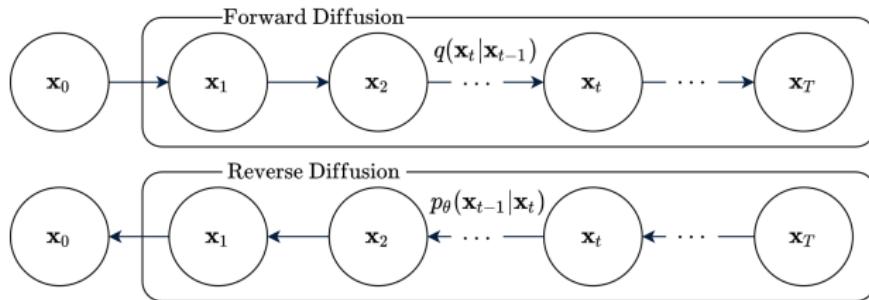
Forward process

1. $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x});$
2. $\mathbf{x}_t = \sqrt{1 - \beta} \mathbf{x}_{t-1} + \sqrt{\beta} \boldsymbol{\epsilon},$
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. $\mathbf{x}_{t-1} = \sigma(\mathbf{x}_t, \theta, t) \cdot \mathbf{x}_t + \mu(\mathbf{x}_t, \theta, t);$
3. $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x});$

Diffusion model



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

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

- ▶ 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

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

Diffusion model

ELBO

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

Statement

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

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

Diffusion model

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

Here

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

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

- ▶ Assume $\sigma^2(\mathbf{x}_t, \theta, t) = \tilde{\beta}_t \mathbf{I}$ (reminder:
 $p(\mathbf{x}_{t-1} | \mathbf{x}_t, \theta) = \mathcal{N}(\mathbf{x}_{t-1} | \mu(\mathbf{x}_t, \theta, t), \sigma^2(\mathbf{x}_t, \theta, t))$).
- ▶ Use KL formula for normal distributions.
- ▶ Use the fact $\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}$

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

Diffusion model

Reparametrization

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

KL term

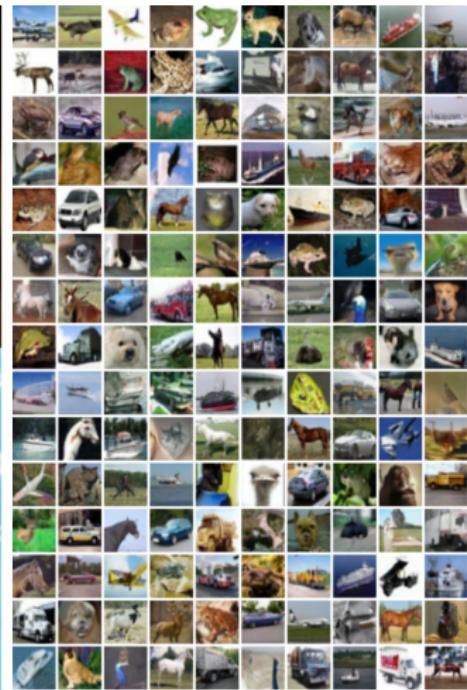
$$\begin{aligned} \mathcal{L}_t &= \mathbb{E}_{\epsilon} \left[\frac{1}{2\tilde{\beta}_t} \left\| \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon \right) - \mu(\mathbf{x}_t, \boldsymbol{\theta}, t) \right\|^2 \right] \\ &= \mathbb{E}_{\epsilon} \left[\frac{\beta_t^2}{2\tilde{\beta}_t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

Noise conditioned score network

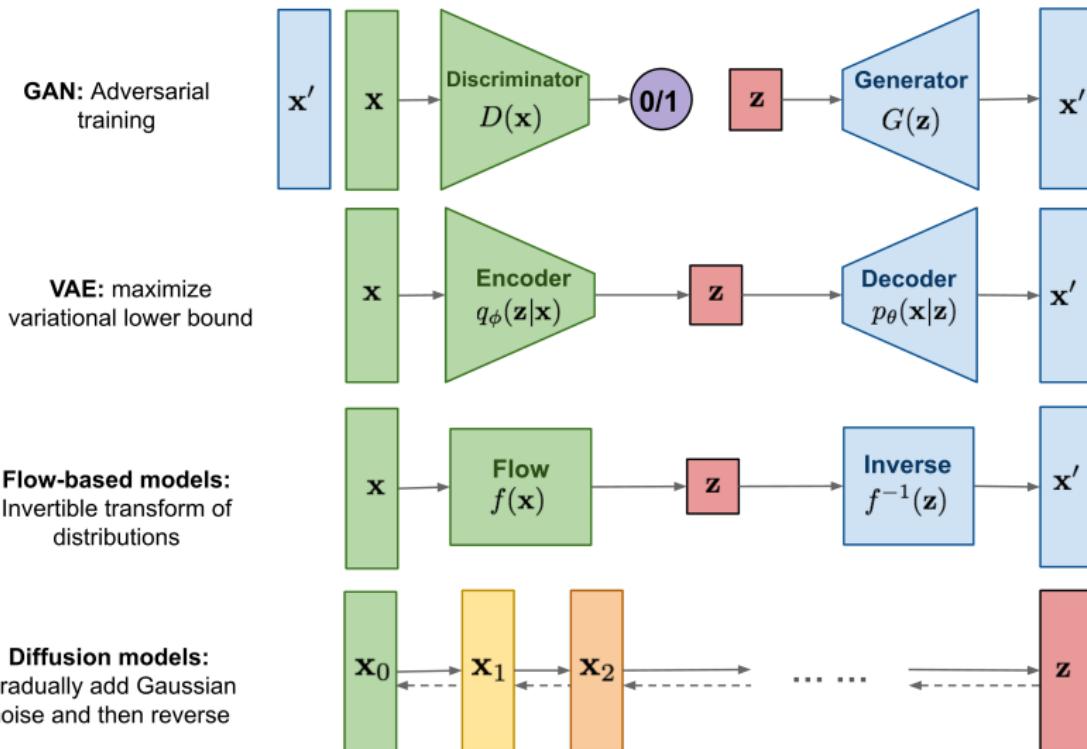
$$\mathbb{E}_{p(\mathbf{x}|\mathbf{x}', \sigma_I)} \|\mathbf{s}(\mathbf{x}, \boldsymbol{\theta}, \sigma_I) - \nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{x}', \sigma_I)\|_2^2 \rightarrow \min_{\boldsymbol{\theta}}$$

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 denoising score matching are two techniques to get scalable algorithm for fitting Fisher divergence.
- ▶ Noise conditioned score nework uses multiple noise levels and annealed Langevin dynamics to fit score function.
- ▶ Gaussian diffusion process is a Markov chain that inject Gaussian noise.
- ▶ Diffusion model is a VAE model which revert gaussian diffusion process using variational inference.
- ▶ Objective of diffusion model is closely related to the noise conditioned score network.