Deep Generative Models

Lecture 13

Roman Isachenko

Moscow Institute of Physics and Technology

2023, Autumn

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

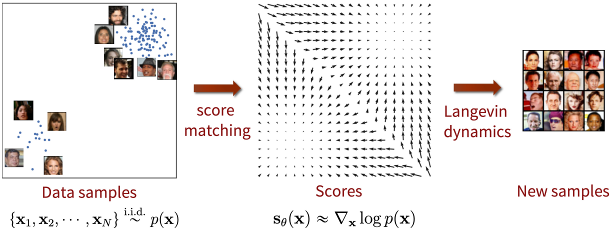
$$\mathbf{x}_{t+1} = \mathbf{x}_t + \eta \frac{1}{2} \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\boldsymbol{\theta}) + \sqrt{\eta} \cdot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}).$$

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

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

Welling M. Bayesian Learning via Stochastic Gradient Langevin Dynamics, 2011

Recap of previous lecture



Theorem (implicit score matching)

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

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

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

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

Theorem (denoising score matching)

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

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

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

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

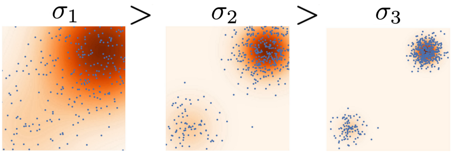
Recap of previous lecture

Noise conditioned score network

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

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

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



Song Y. et al. Generative Modeling by Estimating Gradients of the Data Distribution, 2019

Outline

1. Noise conditioned score network

2. The worst course overview

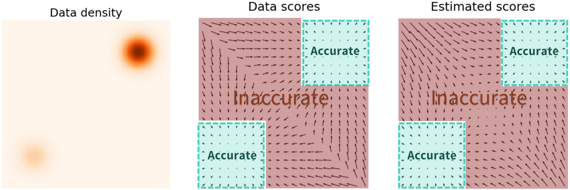
Outline

1. Noise conditioned score network

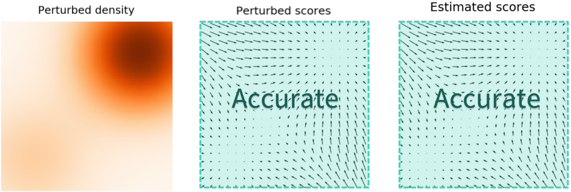
2. The worst course overview

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.



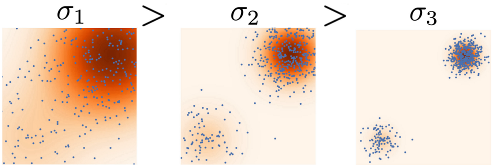
Song Y. Generative Modeling by Estimating Gradients of the Data Distribution, blog post, 2021

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 $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}', \sigma)$ for each noise level:

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

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



Song Y. et al. Generative Modeling by Estimating Gradients of the Data Distribution, 2019

Noise conditioned score network

Training: loss function

$$\sum_{i=1}^{L} \sigma_l^2 \mathbb{E}_{\pi(\mathbf{x})} \mathbb{E}_{\epsilon} \left\| \mathbf{s}_l + \frac{\epsilon}{\sigma_l} \right\|_2^2,$$

Here

- $\mathbf{s}_l = \mathbf{s}_{\theta}(\mathbf{x} + \sigma_l \cdot \epsilon, \sigma_l)$.
- $\nabla_{\mathbf{x}'} \log p(\mathbf{x}'|\mathbf{x}, \sigma) = -\frac{\mathbf{x}' - \mathbf{x}}{\sigma^2} = -\frac{\epsilon}{\sigma_l}$.

Inference: annealed Langevin dynamic

Algorithm 1 Annealed Langevin dynamics.

Require: $\{\sigma_i\}_{i=1}^{L}, \epsilon, T$.

Initialize $\tilde{\mathbf{x}}_0$
for $i \leftarrow 1$ to L **do**
    $\alpha_i \leftarrow \epsilon \cdot \sigma_i^2 / \sigma_L^2$ ▷ α_i is the step size.
    for $t \leftarrow 1$ to T **do**
        Draw $\mathbf{z}_t \sim \mathcal{N}(0, I)$
        $\tilde{\mathbf{x}}_t \leftarrow \tilde{\mathbf{x}}_{t-1} + \frac{\alpha_i}{2} \mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_{t-1}, \sigma_i) + \sqrt{\alpha_i}\, \mathbf{z}_t$
    end for
    $\tilde{\mathbf{x}}_0 \leftarrow \tilde{\mathbf{x}}_T$
end for
return $\tilde{\mathbf{x}}_T$

Samples

Song Y. et al. Improved Techniques for Training Score-Based Generative Models, 2020

Gaussian diffusion model vs Score matching

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

- 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

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

- Let reparametrize our model:

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

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

Outline

1. Noise conditioned score network
2. The worst course overview

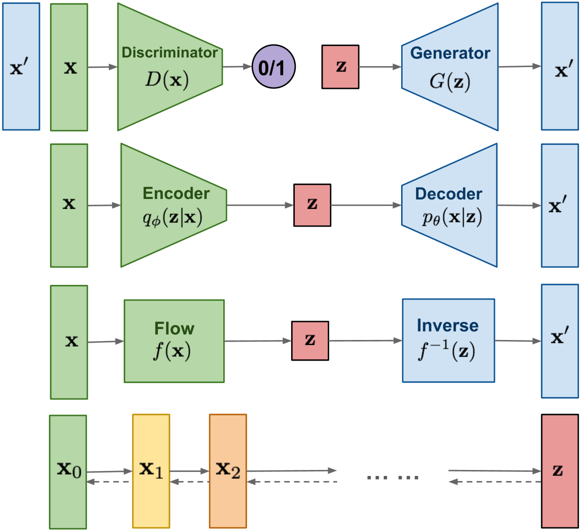
The worst course overview :)

GAN: Adversarial training

VAE: maximize variational lower bound

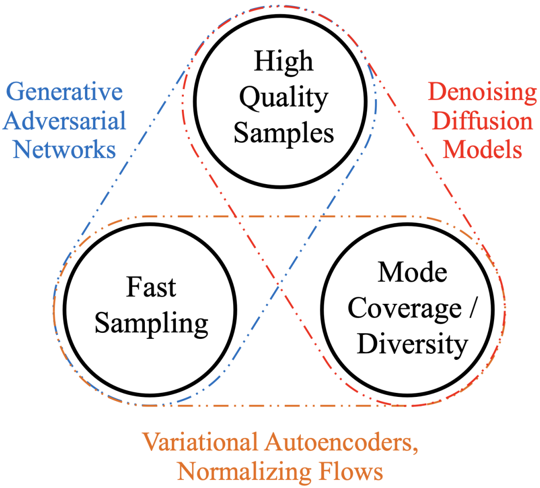
Flow-based models: Invertible transform of distributions

Diffusion models: Gradually add Gaussian noise and then reverse



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

The worst course overview :)



Xiao Z., Kreis K., Vahdat A. Tackling the generative learning trilemma with denoising diffusion GANs, 2021

Summary

- Noise conditioned score network uses multiple noise levels and annealed Langevin dynamics to fit score function.

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