

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)

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

Langevin dynamics

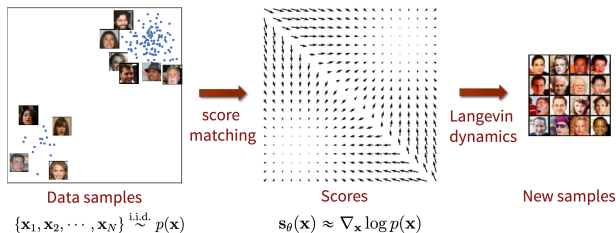
Let \mathbf{x}_0 be a random vector. Then under mild regularity conditions for small enough η samples from the following dynamics

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

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

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

Recap of previous lecture



Theorem (implicit score matching)

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

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

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

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

Theorem (denoising score matching)

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

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

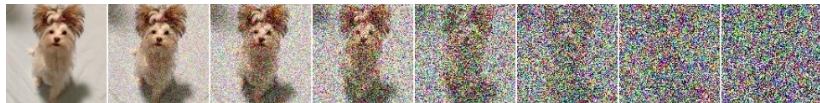
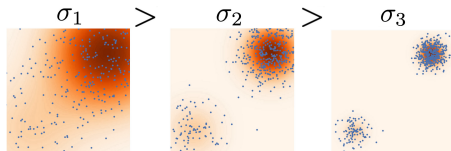
Recap of previous lecture

Noise conditioned score network

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

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

- ▶ 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

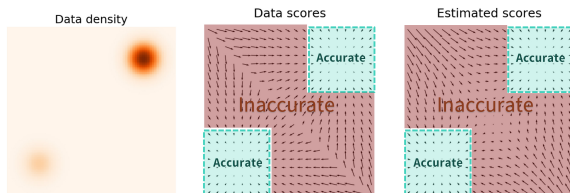
Outline

1. Noise conditioned score network

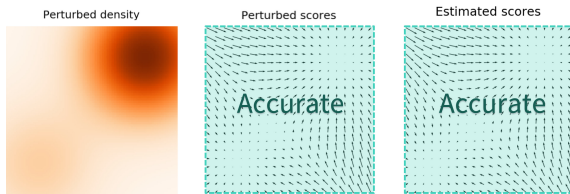
2. The worst course overview

Denoising score matching

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

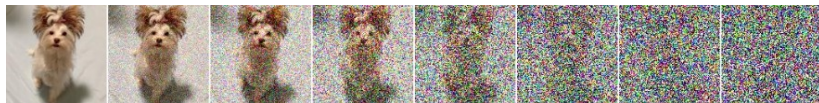
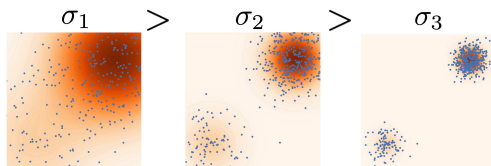


- ▶ If σ is **large**, it is good for low-density regions and multimodal distributions, but we will learn too corrupted distribution.



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 $\mathbf{s}_\theta(\mathbf{x}', \sigma)$ for each noise level:
$$\sum_{l=1}^L \sigma_l^2 \mathbb{E}_{\pi(\mathbf{x})} \mathbb{E}_{p(\mathbf{x}'|\mathbf{x}, \sigma_l)} \|\mathbf{s}_\theta(\mathbf{x}', \sigma_l) - \nabla'_{\mathbf{x}} \log p(\mathbf{x}'|\mathbf{x}, \sigma_l)\|_2^2 \rightarrow \min_{\theta}$$
- ▶ Sample from **annealed** Langevin dynamics (for $l = 1, \dots, L$).



Noise conditioned score network

Training: loss function

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

Here

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

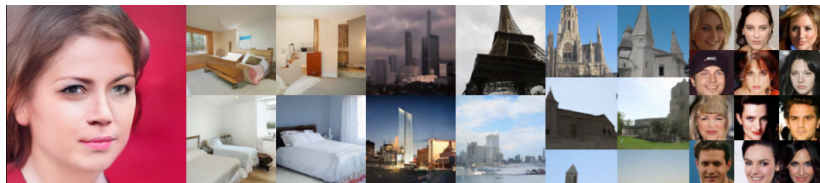
Inference: annealed Langevin dynamic

Algorithm 1 Annealed Langevin dynamics.

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

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

Samples



Gaussian diffusion model vs Score matching

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

- ▶ 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

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

- ▶ Let reparametrize our model:

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

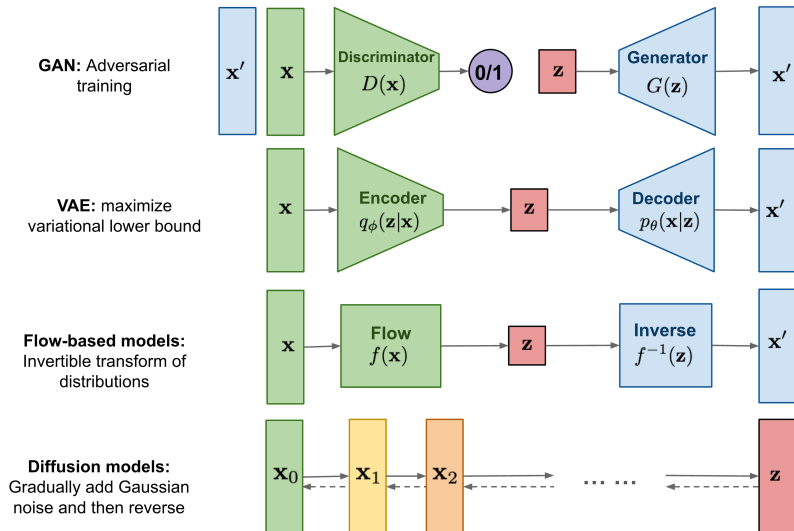
Noise conditioned score network

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

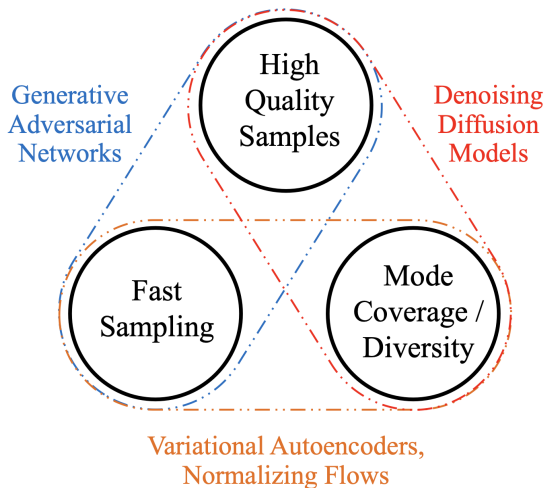
Outline

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

The worst course overview :)



The worst course overview :)



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

Summary

