Deep Generative Models

Lecture 13

Roman Isachenko

Moscow Institute of Physics and Technology

2023, Autumn

Recap of previous lecture

ELBO of gaussian diffusion model

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

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I}),$$
$$p(\mathbf{x}_{t-1}|\mathbf{x}_t, \boldsymbol{\theta}) = \mathcal{N}\big(\mathbf{x}_{t-1}|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}_t, t), \boldsymbol{\sigma}^2_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\big)$$

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

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

Ho J. Denoising Diffusion Probabilistic Models, 2020

Recap of previous lecture

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

Reparametrization

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

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

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

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

Simplified objective

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

Ho J. Denoising Diffusion Probabilistic Models, 2020

Recap of previous lecture

Training

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

Sampling

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

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

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

Ho J. Denoising Diffusion Probabilistic Models, 2020

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

Outline

1. Score matching
   - Implicit score matching
   - Denoising score matching

2. Noise Conditioned Score Network (NCSN)

3. DDPM vs NCSN

Outline

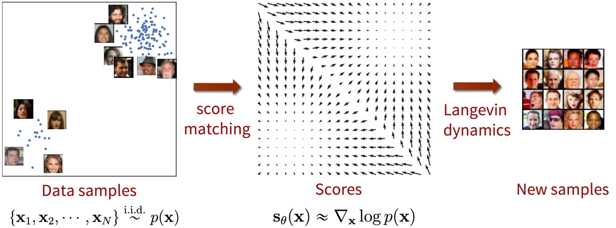
Score matching

We could sample from the model using Langevin dynamics if we have $\nabla_{\mathbf{x}} \log p(\mathbf{x}|\boldsymbol{\theta})$.

Fisher divergence

$$D_F(\pi, p) = \frac{1}{2}\mathbb{E}_\pi \big\| \nabla_{\mathbf{x}} \log p(\mathbf{x}|\boldsymbol{\theta}) - \nabla_{\mathbf{x}} \log \pi(\mathbf{x}) \big\|_2^2 \to \min_{\boldsymbol{\theta}}$$

Let introduce **score function** $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}) = \nabla_{\mathbf{x}} \log p(\mathbf{x}|\boldsymbol{\theta})$.



Problem: we do not know $\nabla_{\mathbf{x}} \log \pi(\mathbf{x})$.

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

Outline

Implicit score matching

Theorem

Under some regularity conditions, it holds

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

Proof (only for 1D)

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

$$\mathbb{E}_\pi[s(x)\nabla_x \log \pi(x)] = \int \pi(x) s(x) \nabla_x \log \pi(x) dx = \int \underbrace{\nabla_x \log p(x)}_{g} \underbrace{\nabla_x \pi(x)}_{\nabla f} dx$$

$$= \underbrace{\nabla_x \log p(x)}_{g} \underbrace{\pi(x)}_{f}\Big|_{-\infty}^{+\infty} - \int \underbrace{\nabla_x \big(\nabla_x \log p(x)\big)}_{\nabla g} \underbrace{\pi(x)}_{f} dx$$

$$= -\mathbb{E}_\pi \nabla_x s(x)$$

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

Hyvarinen A. Estimation of non-normalized statistical models by score matching, 2005

Implicit score matching

Theorem

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

Here $\nabla_{\mathbf{x}}\mathbf{s}_{\boldsymbol\theta}(\mathbf{x}) = \nabla^2_{\mathbf{x}} \log p(\mathbf{x}|\boldsymbol\theta)$ is a Hessian matrix.

1. The right hand side is complex due to Hessian matrix – **sliced score matching**.
2. The left hand side is intractable due to unknown $\pi(\mathbf{x})$ – **denoising score matching**.

Sliced score matching (Hutchinson's trace estimation)

$$\mathrm{tr}\big(\nabla_{\mathbf{x}}\mathbf{s}_{\boldsymbol\theta}(\mathbf{x})\big) = \mathbb{E}_{p(\boldsymbol\epsilon)}\Big[\boldsymbol\epsilon^T \nabla_{\mathbf{x}}\mathbf{s}_{\boldsymbol\theta}(\mathbf{x})\boldsymbol\epsilon\Big]$$

Song Y. Sliced Score Matching: A Scalable Approach to Density and Score Estimation, 2019
Song Y. Generative Modeling by Estimating Gradients of the Data Distribution, blog post, 2021

Outline

Denoising score matching

Let perturb original data $\mathbf{x} \sim \pi(\mathbf{x})$ by random normal noise

$$\mathbf{x}' = \mathbf{x} + \sigma \cdot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \quad p(\mathbf{x}'|\mathbf{x}, \sigma) = \mathcal{N}(\mathbf{x}'|\mathbf{x}, \sigma^2 \mathbf{I})$$

$$\pi(\mathbf{x}'|\sigma) = \int p(\mathbf{x}'|\mathbf{x}, \sigma)\pi(\mathbf{x})d\mathbf{x}.$$

Assumption

The solution of

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

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

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

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

Denoising score matching

Theorem

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

Proof

$$\mathbb{E}_{\pi(\mathbf{x}'|\sigma)} \left\| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}', \sigma) - \nabla_{\mathbf{x}'} \log \pi(\mathbf{x}'|\sigma) \right\|_2^2 =$$
$$= \mathbb{E}_{\pi(\mathbf{x}'|\sigma)} \Big[\| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}', \sigma) \|^2 + \underbrace{\| \nabla_{\mathbf{x}'} \log \pi(\mathbf{x}'|\sigma) \|_2^2}_{\text{const}(\boldsymbol{\theta})} - 2 \mathbf{s}_{\boldsymbol{\theta}}^T(\mathbf{x}', \sigma) \nabla_{\mathbf{x}'} \log \pi(\mathbf{x}'|\sigma) \Big]$$

$$\mathbb{E}_{\pi(\mathbf{x}'|\sigma)} \| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}', \sigma) \|^2 = \int \pi(\mathbf{x}'|\sigma) \| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}', \sigma) \|^2 d\mathbf{x}' =$$
$$= \int \left(\int p(\mathbf{x}'|\mathbf{x}, \sigma) \pi(\mathbf{x}) d\mathbf{x} \right) \| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}', \sigma) \|^2 d\mathbf{x}' = \mathbb{E}_{\pi(\mathbf{x})} \mathbb{E}_{p(\mathbf{x}'|\mathbf{x},\sigma)} \| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}', \sigma) \|^2 d\mathbf{x}'$$

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

Denoising score matching

Theorem

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

Proof (continued)

$$\mathbb{E}_{\pi(\mathbf{x}'|\sigma)}\left[\mathbf{s}_{\boldsymbol{\theta}}^T(\mathbf{x}',\sigma)\nabla_{\mathbf{x}'}\log \pi(\mathbf{x}'|\sigma)\right] = \int \pi(\mathbf{x}'|\sigma)\left[\mathbf{s}_{\boldsymbol{\theta}}^T(\mathbf{x}',\sigma)\frac{\nabla_{\mathbf{x}'}\pi(\mathbf{x}'|\sigma)}{\pi(\mathbf{x}'|\sigma)}\right]d\mathbf{x}' =$$
$$= \int \left[\mathbf{s}_{\boldsymbol{\theta}}^T(\mathbf{x}',\sigma)\nabla_{\mathbf{x}'}\left(\int p(\mathbf{x}'|\mathbf{x},\sigma)\pi(\mathbf{x})d\mathbf{x}\right)\right]d\mathbf{x}' =$$
$$= \int\int \pi(\mathbf{x})\left[\mathbf{s}_{\boldsymbol{\theta}}^T(\mathbf{x}',\sigma)\nabla_{\mathbf{x}'}p(\mathbf{x}'|\mathbf{x},\sigma)\right]d\mathbf{x}'d\mathbf{x} =$$
$$= \int\int \pi(\mathbf{x})p(\mathbf{x}'|\mathbf{x},\sigma)\left[\mathbf{s}_{\boldsymbol{\theta}}^T(\mathbf{x}',\sigma)\nabla_{\mathbf{x}'}\log p(\mathbf{x}'|\mathbf{x},\sigma)\right]d\mathbf{x}'d\mathbf{x} =$$
$$= \mathbb{E}_{\pi(\mathbf{x})}\mathbb{E}_{p(\mathbf{x}'|\mathbf{x},\sigma)}\left[\mathbf{s}_{\boldsymbol{\theta}}^T(\mathbf{x}',\sigma)\nabla_{\mathbf{x}'}\log p(\mathbf{x}'|\mathbf{x},\sigma)\right]$$

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

Denoising score matching

Theorem

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

Proof (continued)

$$\mathbb{E}_{\pi(\mathbf{x}'|\sigma)}\big\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}',\sigma)-\nabla_{\mathbf{x}'}\log\pi(\mathbf{x}'|\sigma)\big\|_2^2=$$
$$=\mathbb{E}_{\pi(\mathbf{x}'|\sigma)}\Big[\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}',\sigma)\|^2-2\mathbf{s}_{\boldsymbol{\theta}}^T(\mathbf{x}',\sigma)\nabla_{\mathbf{x}'}\log\pi(\mathbf{x}'|\sigma)\Big]+\text{const}(\boldsymbol{\theta})=$$
$$=\mathbb{E}_{\pi(\mathbf{x})}\mathbb{E}_{p(\mathbf{x}'|\mathbf{x},\sigma)}\Big[\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}',\sigma)\|^2-2\mathbf{s}_{\boldsymbol{\theta}}^T(\mathbf{x}',\sigma)\nabla_{\mathbf{x}'}\log p(\mathbf{x}'|\mathbf{x},\sigma)\Big]+\text{const}(\boldsymbol{\theta})$$

Gradient of the noise kernel

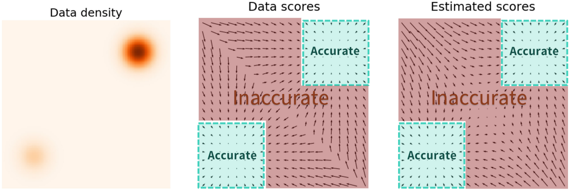
$$\nabla_{\mathbf{x}'}\log p(\mathbf{x}'|\mathbf{x},\sigma)=\nabla_{\mathbf{x}'}\log\mathcal{N}(\mathbf{x}'|\mathbf{x},\sigma^2\mathbf{I})=-\frac{\mathbf{x}'-\mathbf{x}}{\sigma^2}$$

The RHS does not need to compute $\nabla_{\mathbf{x}'}\log\pi(\mathbf{x}'|\sigma)$ and even $\nabla_{\mathbf{x}'}\log\pi(\mathbf{x}')$.

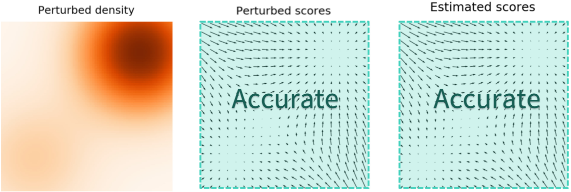
Vincent P. A Connection Between Score Matching and Denoising Autoencoders, 2010

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.



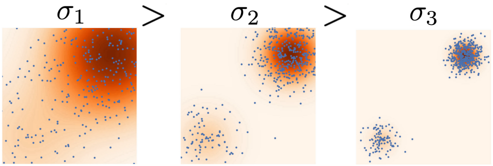
Song Y. Generative Modeling by Estimating Gradients of the Data Distribution, blog post, 2021

Outline

Noise Conditioned Score Network (NCSN)

- Define the sequence of the noise levels: $\sigma_1 > \sigma_2 > \cdots > \sigma_L$.
- Perturb the original data with the different noise levels to obtain $\pi(\mathbf{x}'|\sigma_1), \dots, \pi(\mathbf{x}'|\sigma_L)$.
- Choose σ_1, σ_L such that:

$$\pi(\mathbf{x}'|\sigma_1) \approx \mathcal{N}(0, \sigma_1^2 \mathbf{I}), \quad \pi(\mathbf{x}'|\sigma_L) \approx \pi(\mathbf{x}).$$



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

Noise Conditioned Score Network (NCSN)

Train the denoising score function $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}', \sigma)$ for each noise level using unified weighted objective:

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

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

Training

1. Get the sample $\mathbf{x}_0 \sim \pi(\mathbf{x})$.
2. Sample noise level $l \sim U[1, L]$ and the noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$.
3. Get noisy image $\mathbf{x}' = \mathbf{x}_0 + \sigma_l \cdot \boldsymbol{\epsilon}$.
4. Compute loss $\mathcal{L} = \|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}', \sigma_l) + \frac{\boldsymbol{\epsilon}}{\sigma_l}\|^2$.

How to sample from this model?

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

Noise Conditioned Score Network (NCSN)

Sampling (annealed Langevin dynamics)

- Sample $\mathbf{x}_0 \sim \mathcal{N}(0, \sigma_1 \mathbf{I}) \approx \pi(\mathbf{x}|\sigma_L)$.
- Apply T steps of Langevin dynamic

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \frac{1}{2}\alpha_l \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}, \sigma_l) + \sqrt{\alpha_l}\boldsymbol{\epsilon}_t.$$

- Update $\mathbf{x}_0 := \mathbf{x}_T$ and choose the next σ_l.

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

Outline

DDPM vs NCSN

NCSN objective

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

DDPM objective

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

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

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

Let reparametrize our model:

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

Ho J. Denoising Diffusion Probabilistic Models, 2020

Summary

- Score matching proposes to minimize Fisher divergence to get score function.
- Implicit score matching tries to avoid the value of original distribution $\pi(\mathbf{x})$. Sliced score matching makes implicit score matching scalable.
- Denoising score matching minimizes Fisher divergence on noisy samples.
- 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.