Deep Generative Models

Lecture 13

Roman Isachenko

Ozon Masters

Spring, 2022

Recap of previous lecture

Continuous dynamic

$$\frac{d\mathbf{z}(t)}{dt} = f(\mathbf{z}(t), \boldsymbol{\theta}).$$

Forward pass

$$\mathbf{z}(t_1) = \int_{t_0}^{t_1} f(\mathbf{z}(t), \boldsymbol{\theta}) dt + \mathbf{z}_0 \quad \Rightarrow \quad \text{ODE Solver}$$

Backward pass

$$\left.\begin{aligned} \frac{\partial L}{\partial \boldsymbol{\theta}(t_0)} &= \mathbf{a}_{\boldsymbol{\theta}}(t_0) = -\int_{t_1}^{t_0} \mathbf{a}_{\mathbf{z}}(t)^T \frac{\partial f(\mathbf{z}(t), \boldsymbol{\theta})}{\partial \boldsymbol{\theta}(t)} dt + 0 \\ \frac{\partial L}{\partial \mathbf{z}(t_0)} &= \mathbf{a}_{\mathbf{z}}(t_0) = -\int_{t_1}^{t_0} \mathbf{a}_{\mathbf{z}}(t)^T \frac{\partial f(\mathbf{z}(t), \boldsymbol{\theta})}{\partial \mathbf{z}(t)} dt + \frac{\partial L}{\partial \mathbf{z}(t_1)} \\ \mathbf{z}(t_0) &= -\int_{t_1}^{t_0} f(\mathbf{z}(t), \boldsymbol{\theta}) dt + \mathbf{z}_1. \end{aligned}\right\} \Rightarrow \text{ODE Solver}$$

Chen R. T. Q. et al. Neural Ordinary Differential Equations, 2018

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

$$\log p(\mathbf{z}(t_1)) = \log p(\mathbf{z}(t_0)) - \int_{t_0}^{t_1} \text{tr} \left(\frac{\partial f(\mathbf{z}(t), \boldsymbol{\theta})}{\partial \mathbf{z}(t)} \right) dt =$$

$$= \log p(\mathbf{z}(t_0)) - \mathbb{E}_{p(\boldsymbol{\epsilon})} \int_{t_0}^{t_1} \left[\boldsymbol{\epsilon}^T \frac{\partial f}{\partial \mathbf{z}} \boldsymbol{\epsilon} \right] dt.$$

Grathwohl W. et al. FFJORD: Free-form Continuous Dynamics for Scalable Reversible Generative Models, 2018

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)

$$\mathbf{w}(t) - \mathbf{w}(s) \sim \mathcal{N}(0, t - s), \quad d\mathbf{w} = \boldsymbol{\epsilon} \cdot \sqrt{dt}, \text{ where } \boldsymbol{\epsilon} \sim \mathcal{N}(0, 1).$$

Langevin dynamics

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

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

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

Outline

1. Score matching

Score matching

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

Fisher divergence

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

Score function

$$\mathbf{s}(\mathbf{x}, \boldsymbol{\theta}) = \nabla_{\mathbf{x}} \log p(\mathbf{x}|\boldsymbol{\theta})$$

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

Theorem

Under some regularity conditions, it holds

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

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

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

Score matching

Theorem

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

Proof (only for 1D)

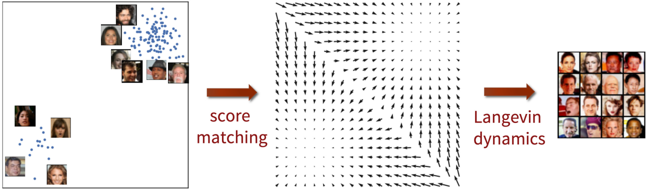
$$\mathbb{E}_{\pi}\left\|s(x)-\nabla_x\log\pi(x)\right\|_2^2=\mathbb{E}_{\pi}\left[s(x)^2+(\nabla_x\log\pi(x))^2-2[s(x)\nabla_x\log\pi(x)]\right]$$

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

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

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

Score matching



Data samples $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \overset{\text{i.i.d.}}{\sim} p(\mathbf{x})$

Scores $\mathbf{s}_\theta(\mathbf{x}) \approx \nabla_\mathbf{x} \log p(\mathbf{x})$

New samples

Theorem

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

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**.

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

Score matching

Sliced score matching (Hutchinson's trace estimation)

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

where $\mathbb{E}[\boldsymbol{\epsilon}] = 0$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = I$.

Denoising score mathing

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

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

satisfies $\mathbf{s}(\mathbf{x},\boldsymbol{\theta},\sigma) \approx \mathbf{s}(\mathbf{x},\boldsymbol{\theta},0) = \mathbf{s}(\mathbf{x},\boldsymbol{\theta})$ using small enough noise scale σ.

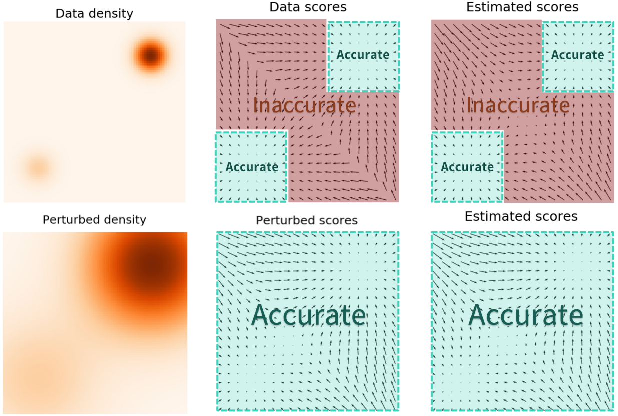
Song Y. Sliced Score Matching: A Scalable Approach to Density and Score Estimation, 2019

Vincent P. A connection between score matching and denoising autoencoders. Neural computation, 2011

Denoising score matching

Theorem

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



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

Summary

- Score matching proposes to minimize Fisher divergence to get score function.
- Sliced score matching and denoised score matching are two techniques to get scalable algorithm for fitting Fisher divergence.