Deep Generative Models

Lecture 10

Roman Isachenko

AI Masters

2024, Spring

Recap of previous lecture

Let take some pretrained image classification model to get the conditional label distribution $p(y|\mathbf{x})$ (e.g. ImageNet classifier).

Evaluation of likelihood-free models

- Sharpness $\Rightarrow$ low $H(y|\mathbf{x}) = -\sum_y \int_{\mathbf{x}} p(y, \mathbf{x}) \log p(y|\mathbf{x}) d\mathbf{x}$.
- Diversity $\Rightarrow$ high $H(y) = -\sum_y p(y) \log p(y)$.

Inception Score

$$IS = \exp(H(y) - H(y|\mathbf{x})) = \exp\left(\mathbb{E}_{\mathbf{x}} KL(p(y|\mathbf{x})||p(y))\right)$$

Frechet Inception Distance

$$D^2(\pi, p) = \|\mathbf{m}_\pi - \mathbf{m}_p\|_2^2 + \mathrm{Tr}\left(\mathbf{\Sigma}_\pi + \mathbf{\Sigma}_p - 2\sqrt{\mathbf{\Sigma}_\pi \mathbf{\Sigma}_p}\right).$$

FID is related to moment matching.

Salimans T. et al. Improved Techniques for Training GANs, 2016
Heusel M. et al. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, 2017

Recap of previous lecture

- $\mathcal{S}_\pi = \{\mathbf{x}_i\}_{i=1}^n \sim \pi(\mathbf{x})$ – real samples;
- $\mathcal{S}_p = \{\mathbf{x}_i\}_{i=1}^n \sim p(\mathbf{x}|\boldsymbol{\theta})$ – generated samples.

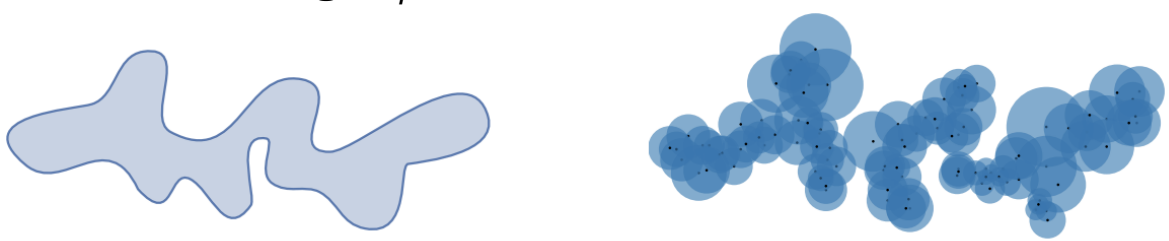
Embed samples using pretrained classifier network (as previously):

$$\mathcal{G}_\pi = \{\mathbf{g}_i\}_{i=1}^n, \quad \mathcal{G}_p = \{\mathbf{g}_i\}_{i=1}^n.$$

Define binary function:

$$f(\mathbf{g}, \mathcal{G}) = \begin{cases} 1, \text{if exists } \mathbf{g}' \in \mathcal{G} : \|\mathbf{g} - \mathbf{g}'\|_2 \leq \|\mathbf{g}' - \mathrm{NN}_k(\mathbf{g}', \mathcal{G})\|_2; \\ 0, \text{otherwise.} \end{cases}$$

$$\text{Precision}(\mathcal{G}_\pi, \mathcal{G}_p) = \frac{1}{n} \sum_{\mathbf{g} \in \mathcal{G}_p} f(\mathbf{g}, \mathcal{G}_\pi); \quad \text{Recall}(\mathcal{G}_\pi, \mathcal{G}_p) = \frac{1}{n} \sum_{\mathbf{g} \in \mathcal{G}_\pi} f(\mathbf{g}, \mathcal{G}_p).$$



(a) True manifold (b) Approx. manifold

Kynkäänniemi T. et al. Improved precision and recall metric for assessing generative models, 2019

Outline

1. Langevin dynamic
2. Denoising score matching
3. Noise Conditioned Score Network (NCSN)

Outline

1. Langevin dynamic
2. Denoising score matching
3. Noise Conditioned Score Network (NCSN)

Langevin dynamic

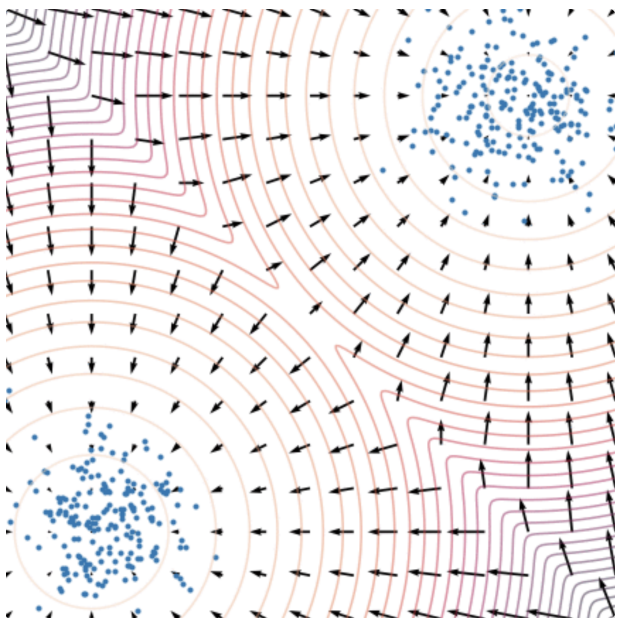
Statement

Let $\mathbf{x}_0$ be a random vector. Then 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 come from $p(\mathbf{x}|\boldsymbol{\theta})$ under mild regularity conditions for small enough η and large enough t.

- Here we assume that we already have some generative model $p(\mathbf{x}|\boldsymbol{\theta})$.
- The density $p(\mathbf{x}|\boldsymbol{\theta})$ is a **stationary** distribution for this SDE.
- What do we get if $\boldsymbol{\epsilon} = \mathbf{0}$?



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

Energy-based models

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

Unnormalized density

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{\hat{p}(\mathbf{x}|\boldsymbol{\theta})}{Z_{\boldsymbol{\theta}}}, \quad \text{where } Z_{\boldsymbol{\theta}} = \int \hat{p}(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x}$$

- $\hat{p}(\mathbf{x}|\boldsymbol{\theta})$ is any non-negative function.
- If we use the reparametrization $\hat{p}(\mathbf{x}|\boldsymbol{\theta}) = \exp(-f_{\boldsymbol{\theta}}(\mathbf{x}))$, we remove the non-negativite constraint.

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}|\boldsymbol{\theta}) = \nabla_{\mathbf{x}} \log \hat{p}(\mathbf{x}|\boldsymbol{\theta}) - \nabla_{\mathbf{x}} \log Z_{\boldsymbol{\theta}} = \nabla_{\mathbf{x}} \log \hat{p}(\mathbf{x}|\boldsymbol{\theta})$$

The gradient of the normalized density equals to the gradient of the unnormalized density.

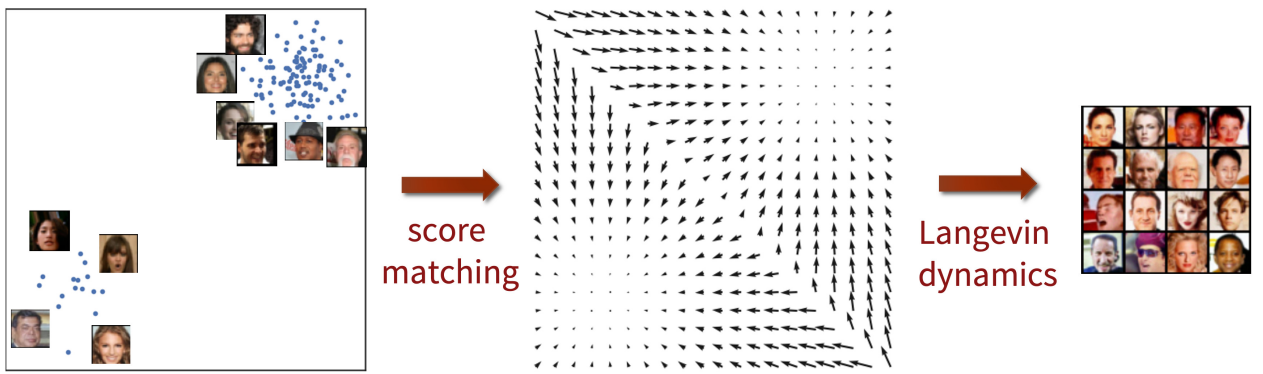
Score matching

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

Fisher divergence

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

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

1. Langevin dynamic
2. Denoising score matching
3. Noise Conditioned Score Network (NCSN)

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

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

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

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

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

Proof (continued)

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

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

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

Gradient of the noise kernel

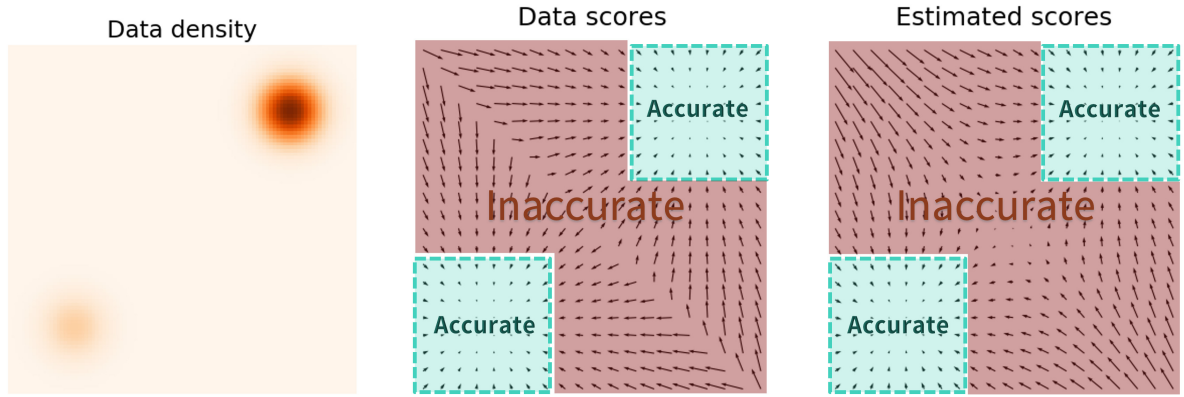
$$\nabla_{\mathbf{x}'} \log p(\mathbf{x}'|\mathbf{x}, \sigma) = \nabla_{\mathbf{x}'} \log \mathcal{N}(\mathbf{x}'|\mathbf{x}, \sigma^2\mathbf{I}) = -\frac{\mathbf{x}' - \mathbf{x}}{\sigma^2}$$

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

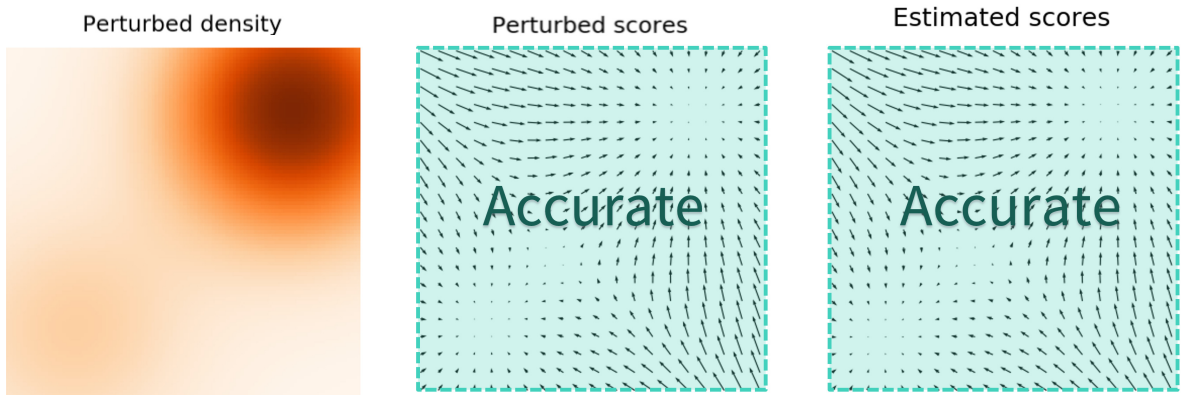
Vincent P. A Connection Between Score Matching and Denoising Autoencoders, 2010

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

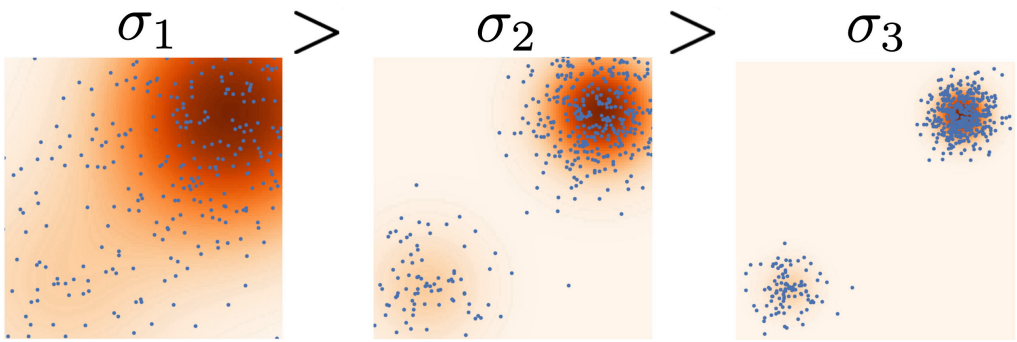
Outline

1. Langevin dynamic
2. Denoising score matching
3. Noise Conditioned Score Network (NCSN)

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 $\pi(\mathbf{x}'|\sigma_1), \ldots, \pi(\mathbf{x}'|\sigma_L)$.
- 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 $\mathbf{x}_0 \sim \mathcal{N}(0, \sigma_1 \mathbf{I}) \approx \pi(\mathbf{x}|\sigma_1)$.
- Apply T steps of Langevin dynamic

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

- 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

Summary

- Langevin dynamics allows to sample from the generative model using the gradient of the log-likelihood.
- Score matching proposes to minimize Fisher divergence to get the score function.
- Denoising score matching minimizes Fisher divergence on noisy samples. It allows to estimate Fisher divergence using samples.
- Noise conditioned score network uses multiple noise levels and annealed Langevin dynamics to fit score function and sample from the model.