Deep Generative Models

Lecture 10

Roman Isachenko

AI Masters

2024, Spring

Recap of previous lecture

How to evaluate likelihood-free models?

$p(y|\mathbf{x})$ – pretrained image classification model (e.g. ImageNet classifier).

What do we want from samples?

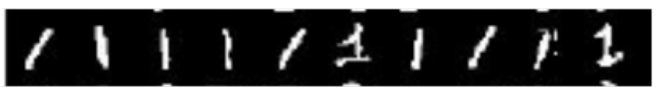
- **Sharpness**

Low sharpness

High sharpness

$p(y|\mathbf{x})$ has low entropy (each image $\mathbf{x}$ should have distinctly recognizable object).

- **Diversity**



Low diversity

High diversity

$p(y) = \int p(y|\mathbf{x})p(\mathbf{x})d\mathbf{x}$ has high entropy (there should be as many classes generated as possible).

---

image credit: https://deepgenerativemodels.github.io

Recap of previous lecture

Frechet Inception Distance (FID)

In case of Normal distributions $\pi(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\pi, \boldsymbol{\Sigma}_\pi)$, $p(\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$:

$$\begin{aligned}\text{FID}(\pi, p) = W_2^2(\pi, p) &= \inf_{\gamma \in \Gamma(\pi, p)} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma} \|\mathbf{x} - \mathbf{y}\|^2 \\ &= \|\boldsymbol{\mu}_\pi - \boldsymbol{\mu}_p\|_2^2 + \text{tr}\left[\boldsymbol{\Sigma}_\pi + \boldsymbol{\Sigma}_p - 2\left(\boldsymbol{\Sigma}_\pi^{1/2}\boldsymbol{\Sigma}_p\boldsymbol{\Sigma}_\pi^{1/2}\right)^{1/2}\right]\end{aligned}$$

Maximum Mean Discrepancy (MMD)

$\pi(\mathbf{x}) = p(\mathbf{y})$ if and only if $\mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}) = \mathbb{E}_{p(\mathbf{y})} f(\mathbf{y})$ for any bounded and continuous f

$$\text{MMD}^2(\pi, p) = \mathbb{E}_{(\mathbf{x},\mathbf{x}')} k(\mathbf{x}, \mathbf{x}') + \mathbb{E}_{(\mathbf{y},\mathbf{y}')} k(\mathbf{y}, \mathbf{y}') - 2\mathbb{E}_{(\mathbf{x},\mathbf{y})} k(\mathbf{x}, \mathbf{y}).$$

Heusel M. et al. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, 2017

Jayasumana S. et al. Rethinking FID: Towards a Better Evaluation Metric for Image Generation, 2024

Recap of previous lecture

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

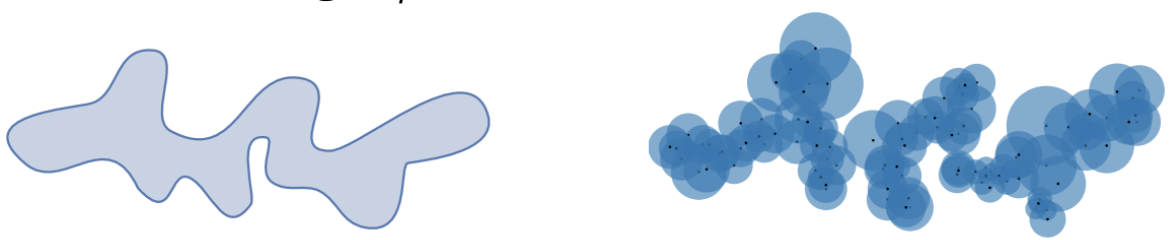
Embed samples using pretrained classifier network (as previously):

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

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

Recap of previous lecture

Langevin dynamic

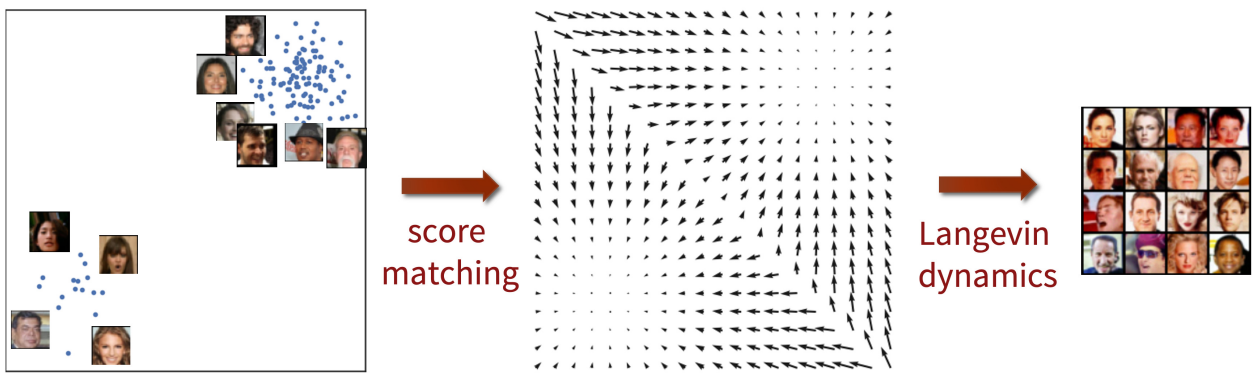
$$\mathbf{x}_{l+1} = \mathbf{x}_l + \frac{\eta}{2} \cdot \nabla_{\mathbf{x}_l} \log p(\mathbf{x}_l|\boldsymbol{\theta}) + \sqrt{\eta} \cdot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}).$$

Fisher divergence

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

Score function

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



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

Outline

1. Denoising score matching

2. Noise Conditioned Score Network (NCSN)

Outline

1. Denoising score matching

2. Noise Conditioned Score Network (NCSN)

Denoising score matching

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

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

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

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

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

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

Proof

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

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

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

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

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

Denoising score matching

Theorem

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

Proof (continued)

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

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

Denoising score matching

Theorem

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

Proof (continued)

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

Gradient of the noise kernel

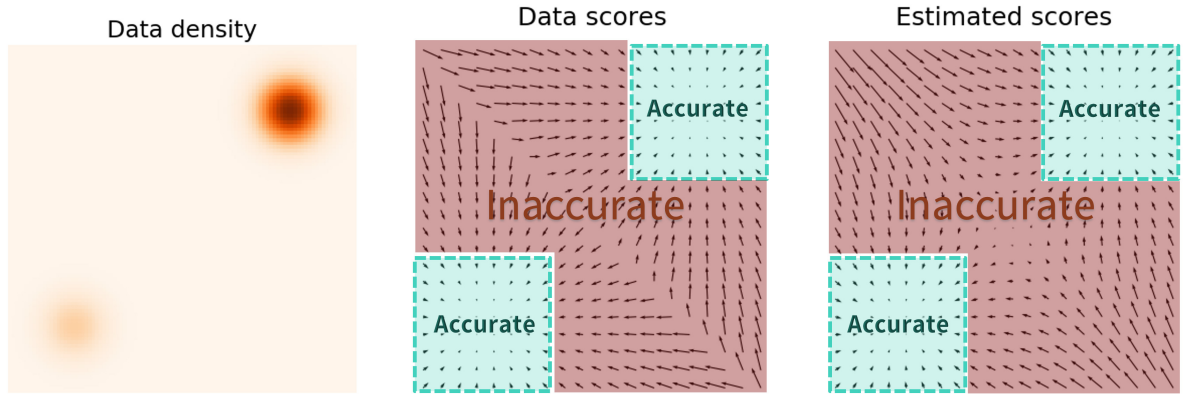
$$\nabla_{\mathbf{x}'} \log q(\mathbf{x}'|\mathbf{x}, \sigma) = \nabla_{\mathbf{x}'} \log \mathcal{N}(\mathbf{x}, \sigma^2\mathbf{I}) = -\frac{\mathbf{x}' - \mathbf{x}}{\sigma^2}$$

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

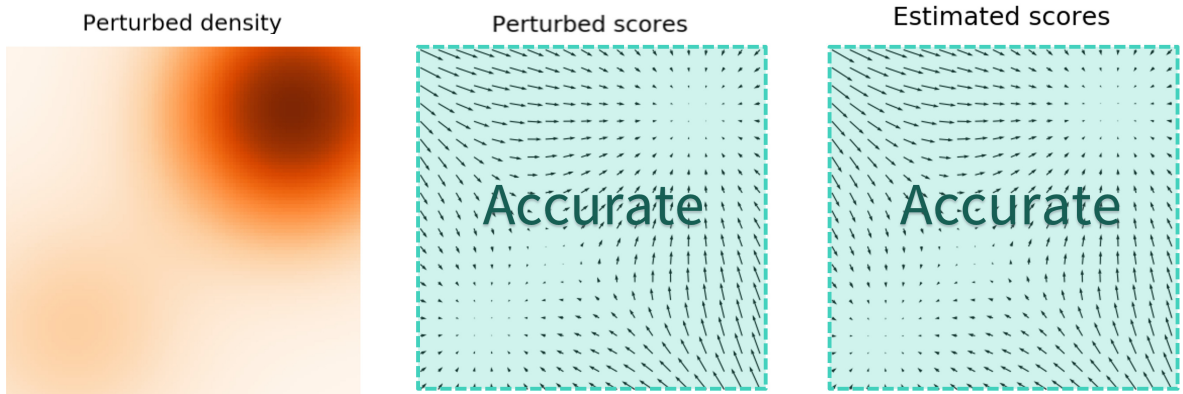
Vincent P. A Connection Between Score Matching and Denoising Autoencoders, 2010

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

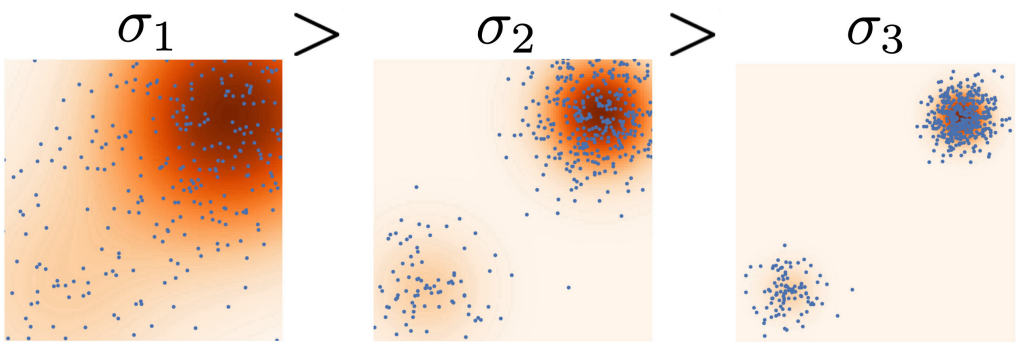
1. Denoising score matching

2. Noise Conditioned Score Network (NCSN)

Noise Conditioned Score Network (NCSN)

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

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



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:

$$\sum_{t=1}^{T} \sigma_t^2 \mathbb{E}_{\pi(\mathbf{x})} \mathbb{E}_{q(\mathbf{x}'|\mathbf{x},\sigma_t)} \left\| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}', \sigma_t) - \nabla_{\mathbf{x}'} \log q(\mathbf{x}'|\mathbf{x}, \sigma_t) \right\|_2^2 \to \min_{\boldsymbol{\theta}}$$

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

Training

1. Get the sample $\mathbf{x}_0 \sim \pi(\mathbf{x})$.
2. Sample noise level $t \sim U[1, T]$ and the noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$.
3. Get noisy image $\mathbf{x}' = \mathbf{x}_0 + \sigma_t \cdot \boldsymbol{\epsilon}$.
4. Compute loss $\mathcal{L} = \|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}', \sigma_t) + \frac{\boldsymbol{\epsilon}}{\sigma_t}\|^2$.

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 L steps of Langevin dynamic

$$\mathbf{x}_l = \mathbf{x}_{l-1} + \frac{\eta_t}{2} \cdot \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_{l-1}, \sigma_t) + \sqrt{\eta_t} \cdot \boldsymbol{\epsilon}_l.$$

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

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

Summary

- 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.