

Deep Generative Models

Lecture 8

Roman Isachenko

Moscow Institute of Physics and Technology
Yandex School of Data Analysis

2025, Autumn

Recap of Previous Lecture

Frechet Inception Distance (FID)

For normal distributions $\pi(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\pi, \boldsymbol{\Sigma}_\pi)$, $p(\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$:

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

- ▶ Requires a large sample size for evaluation
- ▶ FID computation is relatively slow
- ▶ Results are highly dependent on the chosen pretrained classifier
- ▶ Relies on the normality assumption

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

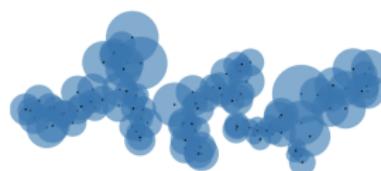
Define a binary indicator function:

$$\mathbb{I}(\mathbf{x}, \mathcal{S}) = \begin{cases} 1, & \text{if } \exists \mathbf{x}' \in \mathcal{S} : \|\mathbf{x} - \mathbf{x}'\|_2 \leq \|\mathbf{x}' - \text{NN}_k(\mathbf{x}', \mathcal{S})\|_2; \\ 0, & \text{otherwise} \end{cases}$$

$$\text{Precision}(\mathcal{S}_\pi, \mathcal{S}_p) = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{S}_\pi} \mathbb{I}(\mathbf{x}, \mathcal{S}_\pi), \quad \text{Recall}(\mathcal{S}_\pi, \mathcal{S}_p) = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{S}_\pi} \mathbb{I}(\mathbf{x}, \mathcal{S}_p)$$



(a) True manifold

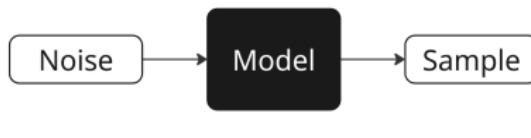


(b) Approx. manifold

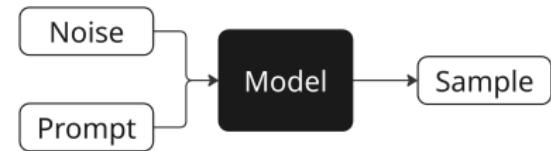
The samples are embedded using a pretrained network, as in FID evaluation.

Recap of Previous Lecture

Unconditional Model

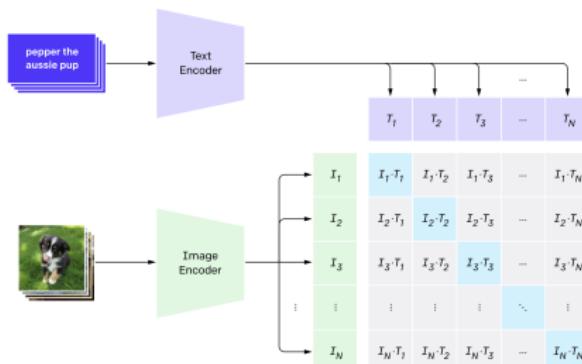


Conditional Model



We require metrics that evaluate not only the quality of generated images, but also their relevance to the prompt.

CLIP Score



Recap of Previous Lecture

- ▶ No perfect automatic evaluation metric exists
- ▶ The most reliable assessment is via human evaluation
- ▶ It's important to evaluate a variety of model aspects

Human Evaluation

Аспект	Yandex ART 2.0	Mj 6.1	Mj 6	Ideogram	Recraft	Google Imagen3	Dall-E 3	FLUX	SBER Kandi3.1
Релевантность	0,59	0,58	0,63	0,45	0,51	0,50	0,50	0,54	0,75
Эстетика	0,49	0,55	0,55	0,51	0,51	0,61	0,61	0,54	0,59
Комплексность	0,44	0,73	0,70	0,68	0,76	0,75	0,75	0,71	0,74
Дефектность	0,69	0,57	0,68	0,55	0,59	0,63	0,63	0,50	0,75
Предпочтение	0,66	0,60	0,69	0,49	0,54	0,63	0,63	0,51	0,84

Recap of Previous Lecture

Langevin Dynamics

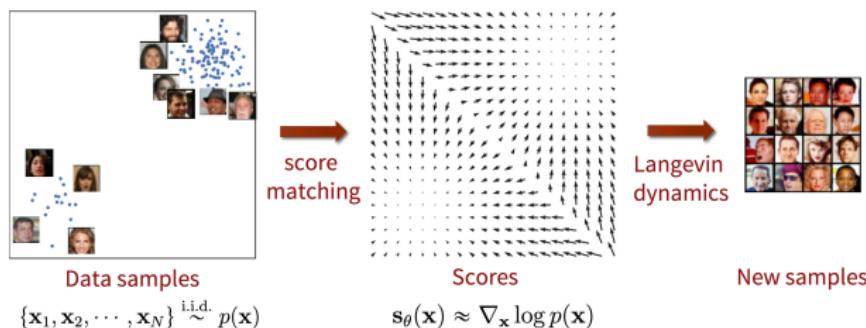
$$\mathbf{x}_{I+1} = \mathbf{x}_I + \frac{\eta}{2} \cdot \nabla_{\mathbf{x}_I} \log p(\mathbf{x}_I | \boldsymbol{\theta}) + \sqrt{\eta} \cdot \boldsymbol{\epsilon}_I, \quad \boldsymbol{\epsilon}_I \sim \mathcal{N}(0, \mathbf{I})$$

Fisher Divergence

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

Score Function

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



Outline

1. Score Matching

Denoising Score Matching

Noise-Conditioned Score Network

2. Forward Gaussian Diffusion Process

3. Denoising Score Matching for Diffusion

4. Reverse Gaussian Diffusion Process

Outline

1. Score Matching

Denoising Score Matching

Noise-Conditioned Score Network

2. Forward Gaussian Diffusion Process

3. Denoising Score Matching for Diffusion

4. Reverse Gaussian Diffusion Process

Outline

1. Score Matching

Denoising Score Matching

Noise-Conditioned Score Network

2. Forward Gaussian Diffusion Process

3. Denoising Score Matching for Diffusion

4. Reverse Gaussian Diffusion Process

Denoising Score Matching

Let us perturb the original data $\mathbf{x} \sim \pi(\mathbf{x})$ by adding Gaussian noise:

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

$$q(\mathbf{x}_\sigma) = \int q(\mathbf{x}_\sigma | \mathbf{x}) \pi(\mathbf{x}) d\mathbf{x}$$

Assumption

The solution to

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

satisfies $\mathbf{s}_{\theta, \sigma}(\mathbf{x}_\sigma) \approx \mathbf{s}_{\theta, 0}(\mathbf{x}_0) = \mathbf{s}_\theta(\mathbf{x})$ if σ is sufficiently small.

- ▶ The score function of the corrupted data closely approximates that of the original data
- ▶ The score function $\mathbf{s}_{\theta, \sigma}(\mathbf{x}_\sigma)$ is parameterized by σ
- ▶ **Note:** Neither $q(\mathbf{x}_\sigma)$ nor $\pi(\mathbf{x})$ are tractable

Denoising Score Matching

Theorem

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

Gradient of the Noise Kernel

$$\mathbf{x}_\sigma = \mathbf{x} + \sigma \epsilon, \quad q(\mathbf{x}_\sigma|\mathbf{x}) = \mathcal{N}(\mathbf{x}, \sigma^2 \mathbf{I})$$

$$\nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma|\mathbf{x}) = -\frac{\mathbf{x}_\sigma - \mathbf{x}}{\sigma^2} = -\frac{\epsilon}{\sigma}$$

- ▶ The right-hand side does not require evaluating $\nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma)$ or $\nabla_{\mathbf{x}_\sigma} \log \pi(\mathbf{x}_\sigma)$
- ▶ The neural network $\mathbf{s}_{\theta,\sigma}(\mathbf{x}_\sigma)$ is trained to **denoise** the corrupted samples \mathbf{x}_σ

Denoising Score Matching

Theorem

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

Proof

$$\begin{aligned}\mathbb{E}_{q(\mathbf{x}_\sigma)} h(\mathbf{x}_\sigma) &= \int q(\mathbf{x}_\sigma) h(\mathbf{x}_\sigma) d\mathbf{x}_\sigma = \\ &= \int \left(\int q(\mathbf{x}_\sigma|\mathbf{x}) \pi(\mathbf{x}) d\mathbf{x} \right) h(\mathbf{x}_\sigma) d\mathbf{x}_\sigma = \mathbb{E}_{\pi(\mathbf{x})} \mathbb{E}_{q(\mathbf{x}_\sigma|\mathbf{x})} h(\mathbf{x}_\sigma)\end{aligned}$$

$$\begin{aligned}\mathbb{E}_{q(\mathbf{x}_\sigma)} \|\mathbf{s}_{\theta,\sigma}(\mathbf{x}_\sigma) - \nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma)\|_2^2 &= \\ &= \mathbb{E}_{q(\mathbf{x}_\sigma)} \left[\|\mathbf{s}_{\theta,\sigma}(\mathbf{x}_\sigma)\|^2 + \underbrace{\|\nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma)\|_2^2}_{\text{const}(\theta)} - 2 \mathbf{s}_{\theta,\sigma}^T(\mathbf{x}_\sigma) \nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma) \right]\end{aligned}$$

Denoising Score Matching

Theorem

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

Proof (Continued)

$$\begin{aligned}\mathbb{E}_{q(\mathbf{x}_\sigma)} [\mathbf{s}_{\theta,\sigma}^T(\mathbf{x}_\sigma) \nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma)] &= \int q(\mathbf{x}_\sigma) \left[\mathbf{s}_{\theta,\sigma}^T(\mathbf{x}_\sigma) \frac{\nabla_{\mathbf{x}_\sigma} q(\mathbf{x}_\sigma)}{q(\mathbf{x}_\sigma)} \right] d\mathbf{x}_\sigma = \\ &= \int \left[\mathbf{s}_{\theta,\sigma}^T(\mathbf{x}_\sigma) \nabla_{\mathbf{x}_\sigma} \left(\int q(\mathbf{x}_\sigma|\mathbf{x}) \pi(\mathbf{x}) d\mathbf{x} \right) \right] d\mathbf{x}_\sigma = \\ &= \int \int \pi(\mathbf{x}) [\mathbf{s}_{\theta,\sigma}^T(\mathbf{x}_\sigma) \nabla_{\mathbf{x}_\sigma} q(\mathbf{x}_\sigma|\mathbf{x})] d\mathbf{x}_\sigma d\mathbf{x} = \\ &= \int \int \pi(\mathbf{x}) q(\mathbf{x}_\sigma|\mathbf{x}) [\mathbf{s}_{\theta,\sigma}^T(\mathbf{x}_\sigma) \nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma|\mathbf{x})] d\mathbf{x}_\sigma d\mathbf{x} = \\ &= \mathbb{E}_{\pi(\mathbf{x})} \mathbb{E}_{q(\mathbf{x}_\sigma|\mathbf{x})} [\mathbf{s}_{\theta,\sigma}^T(\mathbf{x}_\sigma) \nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma|\mathbf{x})]\end{aligned}$$

Denoising Score Matching

Theorem

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

Proof (Continued)

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

Denoising Score Matching

Original objective:

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

Noisy objective:

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

This is equivalent to a denoising task:

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

$$\mathbb{E}_{\pi(\mathbf{x})} \mathbb{E}_{\mathcal{N}(0, \mathbf{I})} \left\| \mathbf{s}_{\theta,\sigma}(\mathbf{x} + \sigma \epsilon) + \frac{\epsilon}{\sigma} \right\|_2^2 \rightarrow \min_{\theta}$$

Langevin Dynamics

$$\mathbf{x}_{l+1} = \mathbf{x}_l + \frac{\eta}{2} \cdot \mathbf{s}_{\theta,\sigma}(\mathbf{x}_l) + \sqrt{\eta} \cdot \epsilon_l, \quad \epsilon_l \sim \mathcal{N}(0, \mathbf{I})$$

Outline

1. Score Matching

Denoising Score Matching

Noise-Conditioned Score Network

2. Forward Gaussian Diffusion Process

3. Denoising Score Matching for Diffusion

4. Reverse Gaussian Diffusion Process

Denoising Score Matching

$$\mathbb{E}_{\pi(\mathbf{x})} \mathbb{E}_{\mathcal{N}(0, \mathbf{I})} \left\| \mathbf{s}_{\theta, \sigma}(\mathbf{x} + \sigma \epsilon) + \frac{\epsilon}{\sigma} \right\|_2^2 \rightarrow \min_{\theta}$$

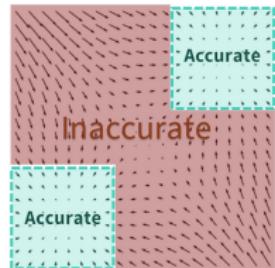
$$\mathbf{x}_{l+1} = \mathbf{x}_l + \frac{\eta}{2} \cdot \mathbf{s}_{\theta, \sigma}(\mathbf{x}_l) + \sqrt{\eta} \cdot \epsilon_l$$

- ▶ For **small** σ , $\mathbf{s}_{\theta, \sigma}(\mathbf{x})$ becomes inaccurate and Langevin dynamics fails to traverse modes
- ▶ For **large** σ , robustness in low-density regions is achieved, but the model learns a distribution that is overly corrupted

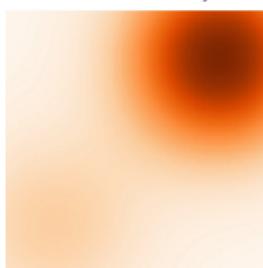
Data density



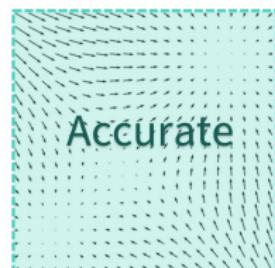
Estimated scores



Perturbed density



Estimated scores

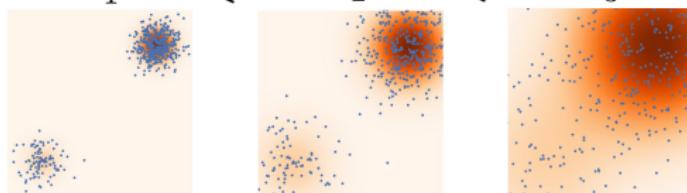


Noise-Conditioned Score Network (NCSN)

- ▶ Specify a sequence of noise levels: $\sigma_1 < \sigma_2 < \dots < \sigma_T$
- ▶ Perturb each data point with different noise levels:
 $\mathbf{x}_t = \mathbf{x} + \sigma_t \epsilon$, so $\mathbf{x}_t \sim q(\mathbf{x}_t)$
- ▶ Choose σ_1, σ_T such that:

$$q(\mathbf{x}_1) \approx \pi(\mathbf{x}), \quad q(\mathbf{x}_T) \approx \mathcal{N}(0, \sigma_T^2 \mathbf{I})$$

$$\sigma_1 \quad < \quad \sigma_2 \quad < \quad \sigma_3$$



Noise-Conditioned Score Network (NCSN)

Train the denoising score function $s_{\theta, \sigma_t}(\mathbf{x}_t)$ for each noise level σ_t using a unified weighted objective:

$$\sum_{t=1}^T \sigma_t^2 \mathbb{E}_{\pi(\mathbf{x})} \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x})} \|s_{\theta, \sigma_t}(\mathbf{x}_t) - \nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t|\mathbf{x})\|_2^2 \rightarrow \min_{\theta}$$

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

Training Procedure

1. Sample $\mathbf{x}_0 \sim \pi(\mathbf{x})$
2. Sample $t \sim U\{1, T\}$ and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
3. Construct noisy image $\mathbf{x}_t = \mathbf{x}_0 + \sigma_t \epsilon$
4. Evaluate loss $\mathcal{L} = \sigma_t^2 \left\| s_{\theta, \sigma_t}(\mathbf{x}_t) + \frac{\epsilon}{\sigma_t} \right\|^2$

How do we sample from such a model?

Noise-Conditioned Score Network (NCSN)

Sampling (Annealed Langevin Dynamics)

- ▶ Sample initial point $\mathbf{x}_0 \sim \mathcal{N}(0, \sigma_T^2 \mathbf{I}) \approx q(\mathbf{x}_T)$
- ▶ At each noise level, apply L steps of Langevin dynamics:

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

- ▶ Update $\mathbf{x}_0 := \mathbf{x}_L$ and reduce to the next lower σ_t



Outline

1. Score Matching

Denoising Score Matching

Noise-Conditioned Score Network

2. Forward Gaussian Diffusion Process

3. Denoising Score Matching for Diffusion

4. Reverse Gaussian Diffusion Process

Forward Gaussian Diffusion Process

Let $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x})$, $\beta_t \ll 1$. Define a Markov chain:

$$\mathbf{x}_t = \sqrt{1 - \beta_t} \mathbf{x}_{t-1} + \sqrt{\beta_t} \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I})$$

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I})$$

Langevin Dynamics

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

$$\begin{aligned} \mathbf{x}_t &= \sqrt{1 - \beta_t} \mathbf{x}_{t-1} + \sqrt{\beta_t} \boldsymbol{\epsilon}_t \approx \left(1 - \frac{\beta_t}{2}\right) \mathbf{x}_{t-1} + \sqrt{\beta_t} \boldsymbol{\epsilon}_t = \\ &\quad \mathbf{x}_{t-1} + \frac{\beta_t}{2} (-\mathbf{x}_{t-1}) + \sqrt{\beta_t} \boldsymbol{\epsilon}_t \end{aligned}$$

- ▶ $\beta_t = \eta$
- ▶ $\nabla_{\mathbf{x}_{t-1}} \log p(\mathbf{x}_{t-1} | \boldsymbol{\theta}) = -\mathbf{x}_{t-1} = \nabla_{\mathbf{x}_{t-1}} \log \mathcal{N}(0, \mathbf{I})$

Forward Gaussian Diffusion Process

$$\mathbf{x}_t = \sqrt{1 - \beta_t} \mathbf{x}_{t-1} + \sqrt{\beta_t} \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I})$$
$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I})$$

Statement 1

Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s = \prod_{s=1}^t (1 - \beta_s)$. Then

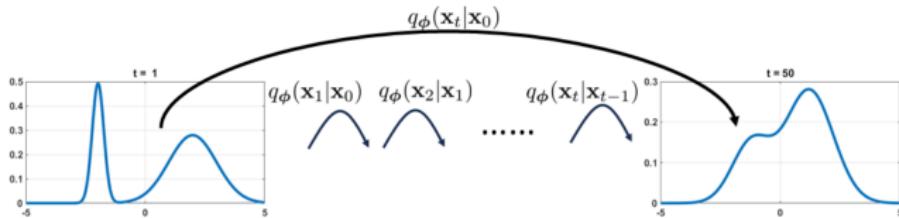
$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I})$$

Thus, samples at any timestep t can be generated directly from \mathbf{x}_0

$$\begin{aligned} \mathbf{x}_t &= \sqrt{\alpha_t} \mathbf{x}_{t-1} + \sqrt{1 - \alpha_t} \boldsymbol{\epsilon}_t = \\ &= \sqrt{\alpha_t} (\sqrt{\alpha_{t-1}} \mathbf{x}_{t-2} + \sqrt{1 - \alpha_{t-1}} \boldsymbol{\epsilon}_{t-1}) + \sqrt{1 - \alpha_t} \boldsymbol{\epsilon}_t = \\ &= \sqrt{\alpha_t \alpha_{t-1}} \mathbf{x}_{t-2} + (\sqrt{\alpha_t (1 - \alpha_{t-1})} \boldsymbol{\epsilon}_{t-1} + \sqrt{1 - \alpha_t} \boldsymbol{\epsilon}_t) = \\ &= \sqrt{\alpha_t \alpha_{t-1}} \mathbf{x}_{t-2} + \sqrt{1 - \alpha_t \alpha_{t-1}} \boldsymbol{\epsilon}'_t \\ &= \dots = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}) \end{aligned}$$

Forward Gaussian Diffusion Process

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N} \left(\sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I} \right); \quad q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N} \left(\sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I} \right)$$



Statement 2

Applying the Markov chain to any distribution $\pi(\mathbf{x})$ yields $\mathbf{x}_\infty \sim p_\infty(\mathbf{x}) = \mathcal{N}(0, \mathbf{I})$, the **stationary** (limiting) distribution:

$$p_\infty(\mathbf{x}) = \int q(\mathbf{x} | \mathbf{x}') p_\infty(\mathbf{x}') d\mathbf{x}'$$

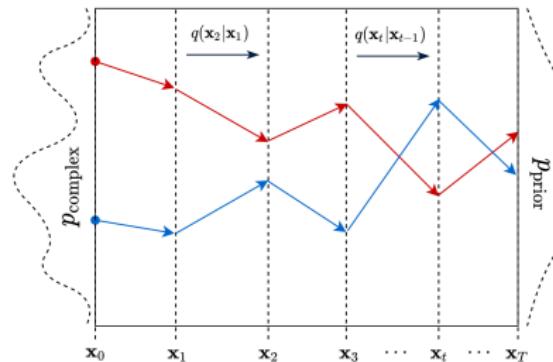
$$p_\infty(\mathbf{x}) = \int q(\mathbf{x}_\infty | \mathbf{x}_0) \pi(\mathbf{x}_0) d\mathbf{x}_0 \approx \mathcal{N}(0, \mathbf{I}) \int \pi(\mathbf{x}_0) d\mathbf{x}_0 = \mathcal{N}(0, \mathbf{I})$$

Chan S. Tutorial on Diffusion Models for Imaging and Vision, 2024

Sohl-Dickstein J. Deep Unsupervised Learning using Nonequilibrium Thermodynamics, 2015

Forward Gaussian Diffusion Process

Diffusion describes the migration of particles from regions of high density to those of low density.



1. $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x})$
2. $\mathbf{x}_t = \sqrt{1 - \beta_t} \mathbf{x}_{t-1} + \sqrt{\beta_t} \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), t \geq 1$
3. After $T \gg 1$ steps: $\mathbf{x}_T \sim p_\infty(\mathbf{x}) = \mathcal{N}(0, \mathbf{I})$

If this process can be reversed, we can sample from $\pi(\mathbf{x})$ by starting from noise $p_\infty(\mathbf{x}) = \mathcal{N}(0, \mathbf{I})$.

Our goal now becomes inverting this diffusion.

Outline

1. Score Matching

Denoising Score Matching

Noise-Conditioned Score Network

2. Forward Gaussian Diffusion Process

3. Denoising Score Matching for Diffusion

4. Reverse Gaussian Diffusion Process

Denoising Score Matching

NCSN

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0, \sigma_t^2 \mathbf{I}), \quad q(\mathbf{x}_1) \approx \pi(\mathbf{x}), \quad q(\mathbf{x}_T) \approx \mathcal{N}(0, \sigma_T^2 \mathbf{I})$$

$$\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t | \mathbf{x}) = -\frac{\mathbf{x}_t - \mathbf{x}}{\sigma_t^2}$$

Gaussian Diffusion

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I}), \quad q(\mathbf{x}_1) \approx \pi(\mathbf{x}), \quad q(\mathbf{x}_T) \approx \mathcal{N}(0, \mathbf{I})$$

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

Theorem (Denoising Score Matching)

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

Note: This enables applying the NCSN approach with annealed Langevin dynamics to diffusion-based denoising models.

Outline

1. Score Matching

Denoising Score Matching

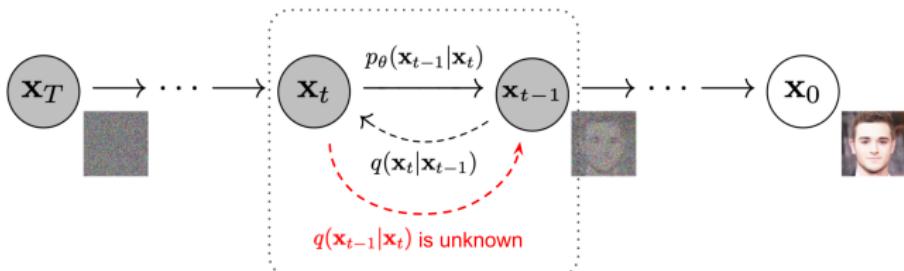
Noise-Conditioned Score Network

2. Forward Gaussian Diffusion Process

3. Denoising Score Matching for Diffusion

4. Reverse Gaussian Diffusion Process

Reverse Gaussian Diffusion Process



Forward Process

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}\left(\sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}\right)$$

Reverse Process

$$q(\mathbf{x}_{t-1} | \mathbf{x}_t) = \frac{q(\mathbf{x}_t | \mathbf{x}_{t-1}) q(\mathbf{x}_{t-1})}{q(\mathbf{x}_t)} \approx p(\mathbf{x}_{t-1} | \mathbf{x}_t, \theta)$$

$q(\mathbf{x}_{t-1})$ and $q(\mathbf{x}_t)$ are intractable:

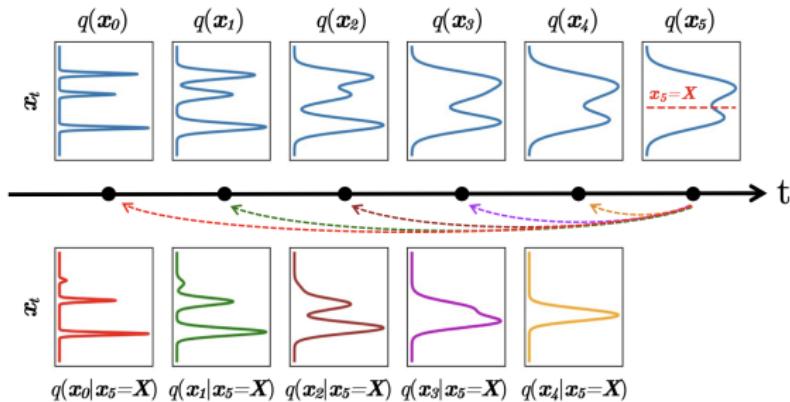
$$q(\mathbf{x}_t) = \int q(\mathbf{x}_t | \mathbf{x}_0) \pi(\mathbf{x}_0) d\mathbf{x}_0$$

Reverse Gaussian Diffusion Process

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t) = \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})q(\mathbf{x}_{t-1})}{q(\mathbf{x}_t)}$$

Theorem (Feller, 1949)

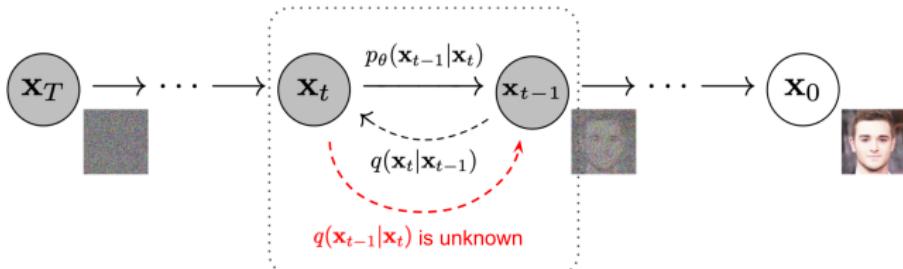
If β_t is sufficiently small, $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is Gaussian (thus, diffusion requires $T \approx 1000$ steps for convergence)



Feller W. On the theory of stochastic processes, with particular reference to applications, 1949

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

Reverse Gaussian Diffusion Process (Ancestral Sampling)



Define the reverse process as:

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t) \approx p(\mathbf{x}_{t-1}|\mathbf{x}_t, \boldsymbol{\theta}) = \mathcal{N}(\mu_{\boldsymbol{\theta},t}(\mathbf{x}_t), \sigma_{\boldsymbol{\theta},t}^2(\mathbf{x}_t))$$

Feller's theorem justifies this Gaussian assumption.

Forward Process

- $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x})$
- $\mathbf{x}_t = \sqrt{1 - \beta_t} \mathbf{x}_{t-1} + \sqrt{\beta_t} \boldsymbol{\epsilon}$
- $\mathbf{x}_T \sim p_{\infty}(\mathbf{x}) = \mathcal{N}(0, \mathbf{I})$

Reverse Process

- $\mathbf{x}_T \sim p_{\infty}(\mathbf{x}) = \mathcal{N}(0, \mathbf{I})$
- $\mathbf{x}_{t-1} = \boldsymbol{\sigma}_{\boldsymbol{\theta},t}(\mathbf{x}_t) \boldsymbol{\epsilon} + \boldsymbol{\mu}_{\boldsymbol{\theta},t}(\mathbf{x}_t)$
- $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x})$

Note: The forward process is non-learnable, i.e., it does not involve trainable parameters

Summary

- ▶ Denoising score matching minimizes the Fisher divergence on corrupted samples, making the divergence estimable via sampling
- ▶ The noise-conditioned score network leverages a range of noise levels and annealed Langevin dynamics to learn the score function and enable sampling
- ▶ The Gaussian diffusion process is a Markov chain that incrementally corrupts data with carefully structured Gaussian noise
- ▶ Denoising score matching, together with Langevin dynamics, can be applied to the Gaussian diffusion process
- ▶ The reverse process reconstructs data from noise samples, although its precise form is intractable