

Deep Generative Models

Lecture 9

Roman Isachenko

Moscow Institute of Physics and Technology
Yandex School of Data Analysis

2024, Autumn

Recap of previous lecture

Langevin dynamic

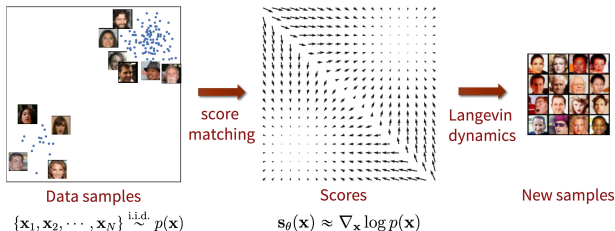
$$\mathbf{x}_{l+1} = \mathbf{x}_l + \frac{\eta}{2} \cdot \nabla_{\mathbf{x}_l} \log p(\mathbf{x}_l | \boldsymbol{\theta}) + \sqrt{\eta} \cdot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}).$$

Fisher divergence

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

Score function

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



Recap of previous lecture

Let perturb original data by normal noise $q(\mathbf{x}_\sigma|\mathbf{x}) = \mathcal{N}(\mathbf{x}, \sigma^2 \cdot \mathbf{I})$

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

Then the solution of

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

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

Theorem (denoising score matching)

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

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

- ▶ We do not need to compute $\nabla_{\mathbf{x}_\sigma} \log q(\mathbf{x}_\sigma)$ at the RHS.
- ▶ $\mathbf{s}_{\theta,\sigma}(\mathbf{x}_\sigma)$ tries to **denoise** a corrupted sample.

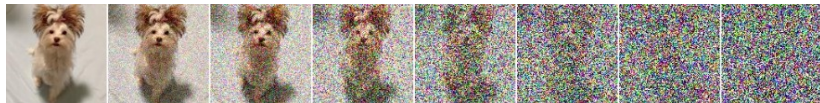
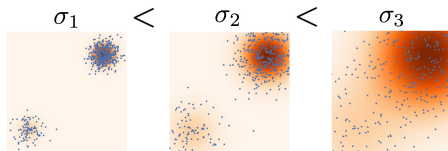
Recap of previous lecture

Noise conditioned score network

- ▶ Define the sequence of noise levels: $\sigma_1 < \sigma_2 < \dots < \sigma_T$.
- ▶ Train denoised score function $\mathbf{s}_{\theta, \sigma_t}(\mathbf{x}_t)$ for each noise level:

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

- ▶ Sample from **annealed** Langevin dynamics (for $t = 1, \dots, T$).



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

Recap of previous lecture

NCSN training

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

NCSN sampling (annealed Langevin dynamics)

- ▶ Sample $\mathbf{x}_0 \sim \mathcal{N}(0, \sigma_T^2 \cdot \mathbf{I}) \approx q(\mathbf{x}_T)$.
- ▶ Apply L steps of Langevin dynamic

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

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

Recap of previous lecture

Forward gaussian diffusion process

Let $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x})$, $\beta_t \in (0, 1)$, $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$.

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

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \cdot \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}(0, \mathbf{I}).$$

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

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

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

If we are able to invert this process, we will get the way to sample $\mathbf{x} \sim \pi(\mathbf{x})$ using noise samples $p_\infty(\mathbf{x}) = \mathcal{N}(0, \mathbf{I})$.

Outline

1. Denoising score matching for diffusion
2. Reverse gaussian diffusion process
3. Gaussian diffusion model as VAE
4. Reparametrization of gaussian diffusion model

Outline

1. Denoising score matching for diffusion
2. Reverse gaussian diffusion process
3. Gaussian diffusion model as VAE
4. Reparametrization of gaussian diffusion model

Denoising score matching

NCSN

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

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

Gaussian diffusion

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

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

Theorem (denoising score matching)

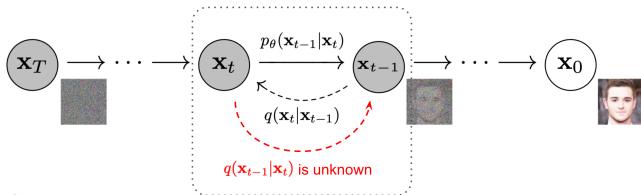
$$\begin{aligned} \mathbb{E}_{q(\mathbf{x}_t)} \left\| \mathbf{s}_{\theta,t}(\mathbf{x}_t) - \nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t) \right\|_2^2 &= \\ &= \mathbb{E}_{\pi(\mathbf{x})} \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x})} \left\| \mathbf{s}_{\theta,t}(\mathbf{x}_t) - \nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t|\mathbf{x})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

Note: We are able to apply NCSN approach with annealed Langevin dynamics to get diffusion denoising model.

Outline

1. Denoising score matching for diffusion
2. Reverse gaussian diffusion process
3. Gaussian diffusion model as VAE
4. Reparametrization of gaussian diffusion model

Reverse gaussian diffusion process



Forward process

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

Reverse process

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

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

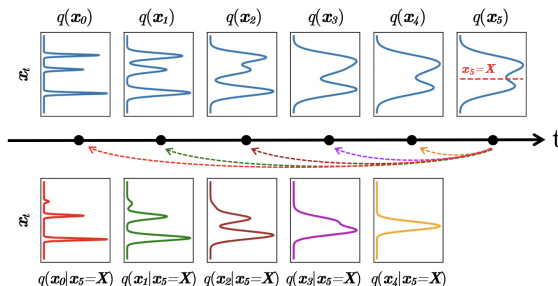
$$q(\mathbf{x}_t) = \int q(\mathbf{x}_t|\mathbf{x}_0)\pi(\mathbf{x}_0)d\mathbf{x}_0$$

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)

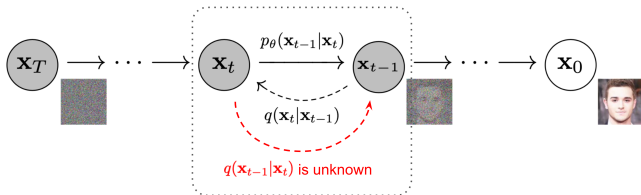
If β_t is small enough, $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ will be Gaussian (that is why diffusion needs $T \approx 1000$ steps to converge).



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



Let define the reverse process:

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

Feller theorem shows that it is a reasonable assumption.

Forward process

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

Reverse process

1. $\mathbf{x}_T \sim p_\infty(\mathbf{x}) = \mathcal{N}(0, \mathbf{I})$;
2. $\mathbf{x}_{t-1} = \sigma_{\theta,t}(\mathbf{x}_t) \cdot \epsilon + \mu_{\theta,t}(\mathbf{x}_t)$;
3. $\mathbf{x}_0 = \mathbf{x} \sim \pi(\mathbf{x})$;

Note: The forward process does not have any learnable parameters!

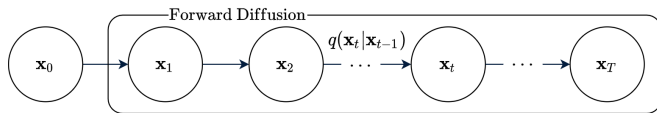
Outline

1. Denoising score matching for diffusion
2. Reverse gaussian diffusion process
3. Gaussian diffusion model as VAE
4. Reparametrization of gaussian diffusion model

Gaussian diffusion model as VAE

Let treat $\mathbf{z} = (\mathbf{x}_1, \dots, \mathbf{x}_T)$ as a latent variable (**note:** each \mathbf{x}_t has the same size). Probabilistic model is

$$p(\mathbf{x}, \mathbf{z} | \theta) = p(\mathbf{x} | \mathbf{z}, \theta) p(\mathbf{z} | \theta)$$



Forward diffusion

- Variational posterior distribution (encoder)

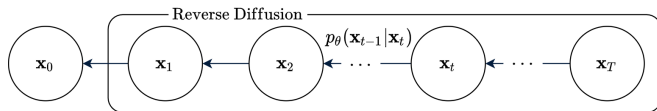
$$q(\mathbf{z} | \mathbf{x}) = q(\mathbf{x}_1, \dots, \mathbf{x}_T | \mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t | \mathbf{x}_{t-1}).$$

- **Note:** there is no learnable parameters.

Gaussian diffusion model as VAE

Let treat $\mathbf{z} = (\mathbf{x}_1, \dots, \mathbf{x}_T)$ as a latent variable (**note:** each \mathbf{x}_t has the same size). Probabilistic model is

$$p(\mathbf{x}, \mathbf{z} | \boldsymbol{\theta}) = p(\mathbf{x} | \mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z} | \boldsymbol{\theta})$$



Reverse diffusion

- Generative distribution (decoder)

$$p(\mathbf{x} | \mathbf{z}, \boldsymbol{\theta}) = p(\mathbf{x}_0 | \mathbf{x}_1, \boldsymbol{\theta}).$$

- Prior distribution

$$p(\mathbf{z} | \boldsymbol{\theta}) = \prod_{t=2}^T p(\mathbf{x}_{t-1} | \mathbf{x}_t, \boldsymbol{\theta}) \cdot p(\mathbf{x}_T).$$

Conditioned reverse distribution

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

$$\begin{aligned} q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) &= \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \\ &= \frac{\mathcal{N}(\sqrt{1-\beta_t} \cdot \mathbf{x}_{t-1}, \beta_t \mathbf{I}) \cdot \mathcal{N}(\sqrt{\bar{\alpha}_{t-1}} \cdot \mathbf{x}_0, (1-\bar{\alpha}_{t-1}) \cdot \mathbf{I})}{\mathcal{N}(\sqrt{\bar{\alpha}_t} \cdot \mathbf{x}_0, (1-\bar{\alpha}_t) \cdot \mathbf{I})} \\ &= \mathcal{N}(\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \cdot \mathbf{I}) \end{aligned}$$

$$\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} \cdot \mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t} \cdot \mathbf{x}_0;$$

$$\tilde{\beta}_t = \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} = \text{const.}$$

$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ defines how to denoise a noisy image \mathbf{x}_t with access to what the final, completely denoised image \mathbf{x}_0 should be.

ELBO for gaussian diffusion model

Standard ELBO

$$\log p(\mathbf{x}|\boldsymbol{\theta}) \geq \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{x})} = \mathcal{L}_{\phi, \boldsymbol{\theta}}(\mathbf{x}) \rightarrow \max_{q, \boldsymbol{\theta}}$$

Derivation

$$\begin{aligned}\mathcal{L}_{\phi, \boldsymbol{\theta}}(\mathbf{x}) &= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \log \frac{p(\mathbf{x}_0, \mathbf{x}_{1:T}|\boldsymbol{\theta})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \\&= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \log \frac{p(\mathbf{x}_T) \prod_{t=1}^T p(\mathbf{x}_{t-1}|\mathbf{x}_t, \boldsymbol{\theta})}{\prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1})} \\&= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \log \frac{p(\mathbf{x}_T) \prod_{t=1}^T p(\mathbf{x}_{t-1}|\mathbf{x}_t, \boldsymbol{\theta})}{\prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)} \\&= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \log \frac{p(\mathbf{x}_T) p(\mathbf{x}_0|\mathbf{x}_1, \boldsymbol{\theta}) \prod_{t=2}^T p(\mathbf{x}_{t-1}|\mathbf{x}_t, \boldsymbol{\theta})}{q(\mathbf{x}_1|\mathbf{x}_0) \prod_{t=2}^T q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)}\end{aligned}$$

We add conditioning on \mathbf{x}_0 to make reverse distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ tractable and to get KL divergences.

ELBO for gaussian diffusion model

Derivation (continued)

$$\begin{aligned}\mathcal{L}_{\phi, \theta}(\mathbf{x}) &= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \log \frac{p(\mathbf{x}_T)p(\mathbf{x}_0|\mathbf{x}_1, \theta) \prod_{t=2}^T p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta)}{q(\mathbf{x}_1|\mathbf{x}_0) \prod_{t=2}^T q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)} = \\&= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \log \frac{p(\mathbf{x}_T)p(\mathbf{x}_0|\mathbf{x}_1, \theta) \prod_{t=2}^T p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta)}{q(\mathbf{x}_1|\mathbf{x}_0) \prod_{t=2}^T \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}} = \\&= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \log \frac{p(\mathbf{x}_T)p(\mathbf{x}_0|\mathbf{x}_1, \theta) \prod_{t=2}^T p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta)}{q(\mathbf{x}_T|\mathbf{x}_0) \prod_{t=2}^T q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} = \\&= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log p(\mathbf{x}_0|\mathbf{x}_1, \theta) + \log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)} + \sum_{t=2}^T \log \left(\frac{p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} \right) \right] = \\&= \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)} \log p(\mathbf{x}_0|\mathbf{x}_1, \theta) + \mathbb{E}_{q(\mathbf{x}_T|\mathbf{x}_0)} \log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)} + \\&\quad + \sum_{t=2}^T \mathbb{E}_{q(\mathbf{x}_{t-1}, \mathbf{x}_t|\mathbf{x}_0)} \log \left(\frac{p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} \right)\end{aligned}$$

ELBO for gaussian diffusion model

$$\begin{aligned}\mathcal{L}_{\phi, \theta}(\mathbf{x}) &= \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)} \log p(\mathbf{x}_0|\mathbf{x}_1, \theta) + \mathbb{E}_{q(\mathbf{x}_T|\mathbf{x}_0)} \log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)} + \\ &\quad + \sum_{t=2}^T \mathbb{E}_{q(\mathbf{x}_{t-1}, \mathbf{x}_t|\mathbf{x}_0)} \log \left(\frac{p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} \right) = \\ &= \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)} \log p(\mathbf{x}_0|\mathbf{x}_1, \theta) - \textcolor{violet}{KL}(q(\mathbf{x}_T|\mathbf{x}_0) || p(\mathbf{x}_T)) - \\ &\quad - \sum_{t=2}^T \underbrace{\mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} \textcolor{violet}{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) || p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta))}_{\mathcal{L}_t}\end{aligned}$$

- **First term** is a decoder distribution

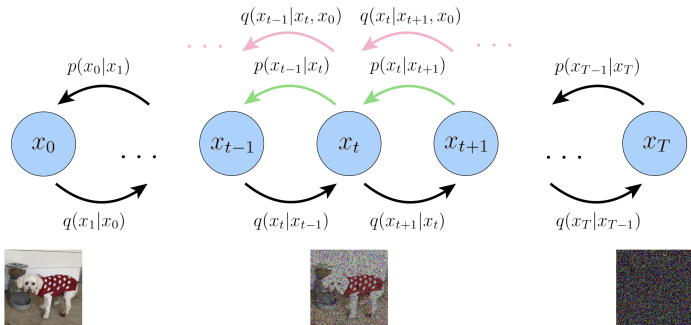
$$\log p(\mathbf{x}_0|\mathbf{x}_1, \theta) = \log \mathcal{N}(\mathbf{x}_0 | \mu_{\theta, t}(\mathbf{x}_1), \sigma_{\theta, t}^2(\mathbf{x}_1)).$$

- **Second term** is constant ($p(\mathbf{x}_T)$ is a standard Normal, $q(\mathbf{x}_T|\mathbf{x}_0)$ is a non-parametrical Normal).

ELBO for gaussian diffusion model

$$\mathcal{L}_{\phi, \theta}(\mathbf{x}) = \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)} \log p(\mathbf{x}_0|\mathbf{x}_1, \theta) - \textcolor{violet}{KL}(q(\mathbf{x}_T|\mathbf{x}_0)||p(\mathbf{x}_T)) -$$

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



Outline

1. Denoising score matching for diffusion
2. Reverse gaussian diffusion process
3. Gaussian diffusion model as VAE
4. Reparametrization of gaussian diffusion model

Reparametrization of DDPM

$$\mathcal{L}_t = \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} KL(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) || p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta))$$

\mathcal{L}_t is the mean of KL between two normal distributions:

$$\begin{aligned} q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) &= \mathcal{N}(\mathbf{x}_{t-1} | \tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I}), \\ p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta) &= \mathcal{N}(\mathbf{x}_{t-1} | \mu_{\theta,t}(\mathbf{x}_t), \sigma_{\theta,t}^2(\mathbf{x}_t)) \end{aligned}$$

Here $\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0)$, $\tilde{\beta}_t = \frac{\beta_t(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}$ have analytical expressions. Let assume

$$\sigma_{\theta,t}^2(\mathbf{x}_t) = \tilde{\beta}_t \mathbf{I} \quad \Rightarrow \quad p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta) = \mathcal{N}(\mathbf{x}_{t-1} | \mu_{\theta,t}(\mathbf{x}_t), \tilde{\beta}_t \mathbf{I}).$$

Theoretically optimal $\sigma_{\theta,t}^2(\mathbf{x}_t)$ lies in the range $[\tilde{\beta}_t, \beta_t]$:

- ▶ β_t is optimal for $\mathbf{x}_0 \sim \mathcal{N}(0, \mathbf{I})$;
- ▶ $\tilde{\beta}_t$ is optimal for $\mathbf{x}_0 \sim \delta(\mathbf{x}_0 - \mathbf{x}^*)$.

Reparametrization of DDPM

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I});$$

$$p(\mathbf{x}_{t-1}|\mathbf{x}_t, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}_{t-1}|\boldsymbol{\mu}_{\boldsymbol{\theta},t}(\mathbf{x}_t), \tilde{\beta}_t \mathbf{I}).$$

Use the formula for KL between two normal distributions:

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

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \cdot \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \cdot \boldsymbol{\epsilon} \quad \Rightarrow \quad \mathbf{x}_0 = \frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \cdot \boldsymbol{\epsilon}}{\sqrt{\bar{\alpha}_t}}$$

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) &= \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \cdot \mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)}{1 - \bar{\alpha}_t} \cdot \mathbf{x}_0 \\ &= \frac{1}{\sqrt{\alpha_t}} \cdot \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{\alpha_t}(1 - \bar{\alpha}_t)} \cdot \boldsymbol{\epsilon}\end{aligned}$$

Reparametrization of DDPM

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

Reparametrization

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

$$\mu_{\theta,t}(\mathbf{x}_t) = \frac{1}{\sqrt{\alpha_t}} \cdot \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{\alpha_t(1 - \bar{\alpha}_t)}} \cdot \epsilon_{\theta,t}(\mathbf{x}_t)$$

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

At each step of reverse diffusion process we try to predict the noise ϵ that we used in the forward diffusion process!

Reparametrization of DDPM

$$\begin{aligned}\mathcal{L}_{\phi, \theta}(\mathbf{x}) &= \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)} \log p(\mathbf{x}_0|\mathbf{x}_1, \theta) - \text{KL}(q(\mathbf{x}_T|\mathbf{x}_0) || p(\mathbf{x}_T)) - \\ &\quad - \sum_{t=2}^T \underbrace{\mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} \text{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) || p(\mathbf{x}_{t-1}|\mathbf{x}_t, \theta))}_{\mathcal{L}_t} \\ \mathcal{L}_t &= \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \mathbf{I})} \left[\frac{(1 - \alpha_t)^2}{2\tilde{\beta}_t\alpha_t(1 - \bar{\alpha}_t)} \left\| \epsilon - \epsilon_{\theta, t}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon) \right\|^2 \right]\end{aligned}$$

Simplified objective

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

Summary

- ▶ Reverse process allows to sample from the real distribution $\pi(\mathbf{x})$ using samples from noise, but it is intractable.
- ▶ We will use approximation to get the reverse process.
- ▶ Diffusion model is a VAE model which reverts gaussian diffusion process using variational inference.
- ▶ ELBO of DDPM could be represented as a sum of KL terms.
- ▶ DDPM is a VAE model with hierarchical latent variables.
- ▶ At each step DDPM predicts the noise that was used in the forward diffusion process.