Lecture 17

Diffusion Models

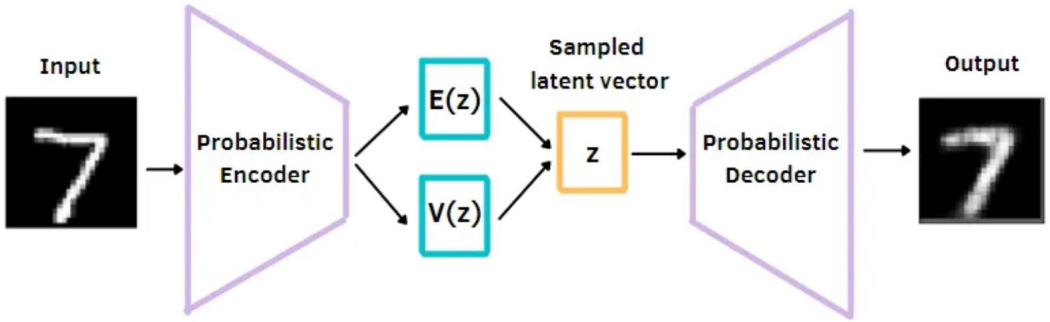
This presentation draws upon insights and information from various sources, including Kevin P. Murphy's 'Probabilistic Machine Learning: Advanced Topics,' Lilian Weng's 'What are Diffusion Models?', 'Denoising Diffusion Probabilistic Models' by Jonathan Ho, Ajay Jain, and Pieter Abbeel, and the tutorial on Denoising Diffusion Probabilistic Models by Jiaming Song, Chenlin Meng, and Arash Vahdat. The content presented is a synthesis of these resources, integrated and interpreted through my own understanding and analysis.

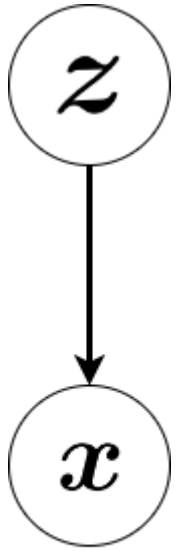
Diffusion Models

- Diffusion models take their inspiration from the principles of non-equilibrium thermodynamics

- Diffusion models use a Markov chain to gradually introduce noise into data, then learn to reverse this process to recreate the desired data from the noise.

Variational Autoencoder





VAE

Image: Eugenia Anello

Diffusion Models

$$x_0 \longrightarrow x_1 \longrightarrow \cdots \longrightarrow x_T$$

Data

Noise

Image: Arash Vahdat

Diffusion Models

$$x_0 \longrightarrow x_1 \longrightarrow \cdots \longrightarrow x_T$$

Data

Noise

$$x_0 \longleftarrow x_1 \longleftarrow \cdots \longleftarrow x_T$$

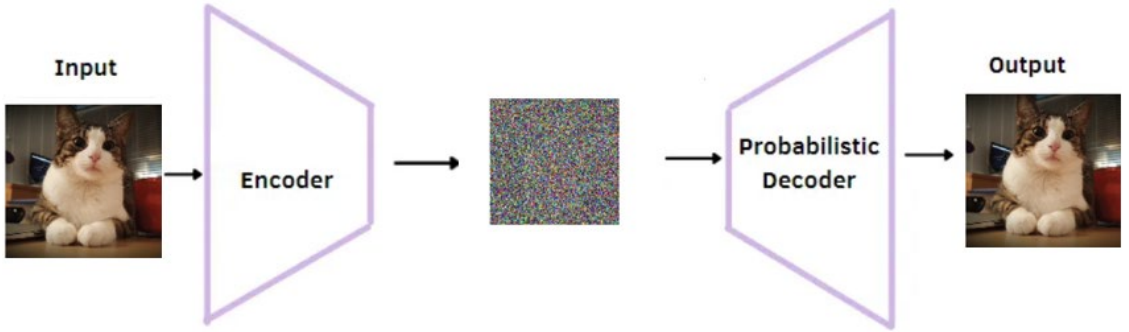
Image: Arash Vahdat

Diffusion Models vs VAE

$$x_0 \longrightarrow x_1 \longrightarrow \cdots \longrightarrow x_T$$

Data  Noise

$$x_0 \longleftarrow x_1 \longleftarrow \cdots \longleftarrow x_T$$



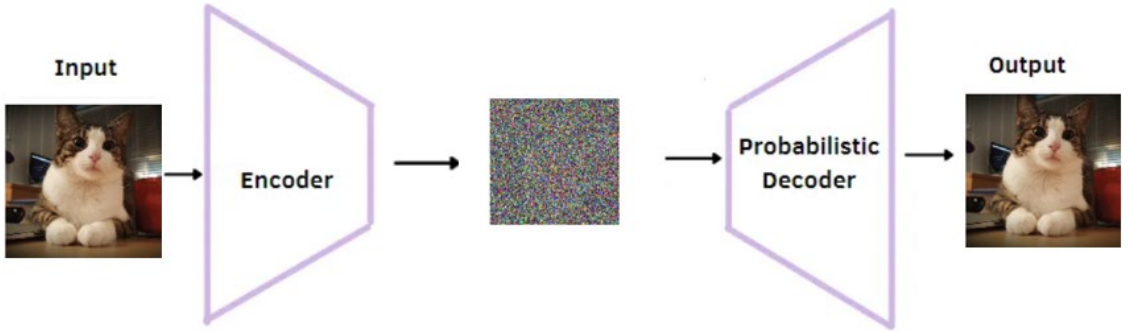
Diffusion Models vs VAE

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

Data

Noise

$$x_0 \longleftarrow x_1 \longleftarrow \cdots \longleftarrow x_T$$



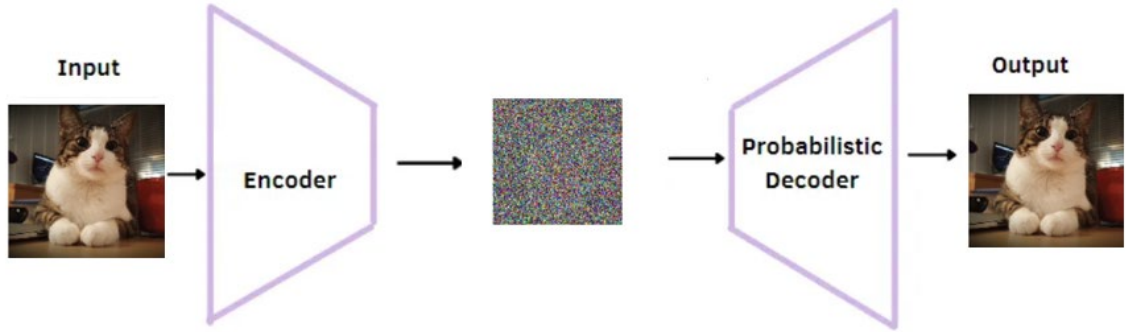
Diffusion Models vs VAE

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

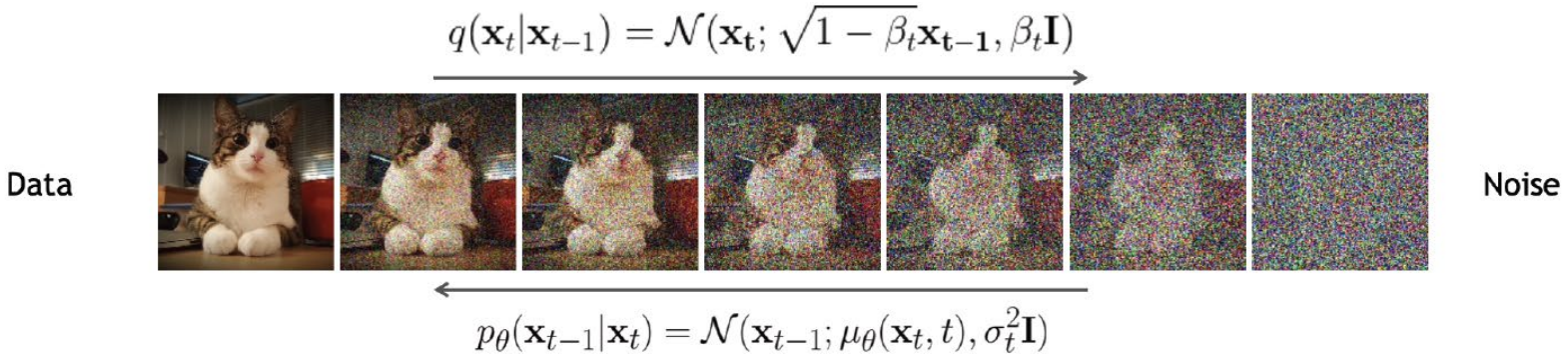
Data

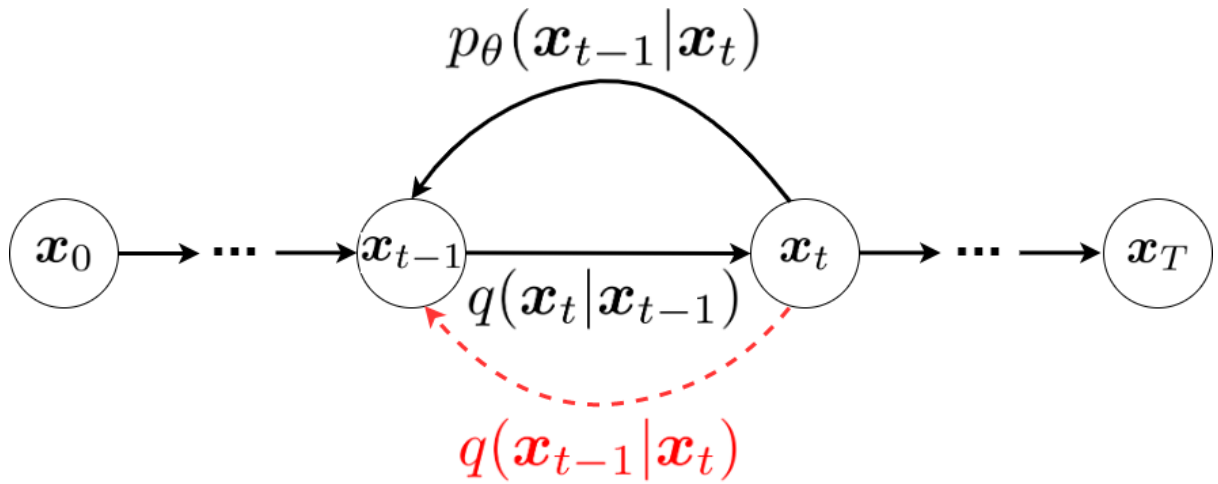
Noise

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I})$$

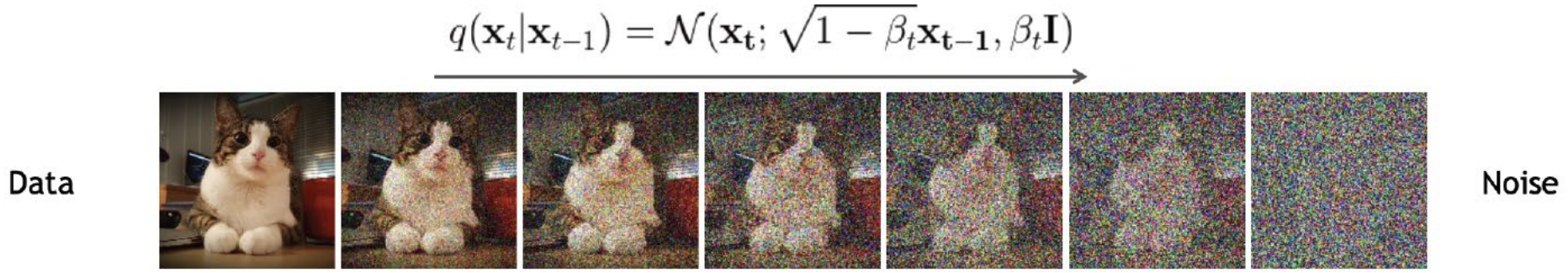


Denoising Diffusion Probabilistic Models (DDPMs)





The Forwards Encoder Process in DDPMs



The forward process is defined by the equation:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t|\sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

Where:

- β_t is a noise level parameter, part of a predetermined noise schedule.

$$\beta_1 < \beta_2 < \cdots < \beta_T$$

$$\beta_t \in (0, 1)$$

Joint Distribution over Latent States in DDPMs

The joint distribution is expressed as:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$$

Where:

- $x_{1:T}$ represents all latent states from time 1 to T.
- x_0 is the initial state.

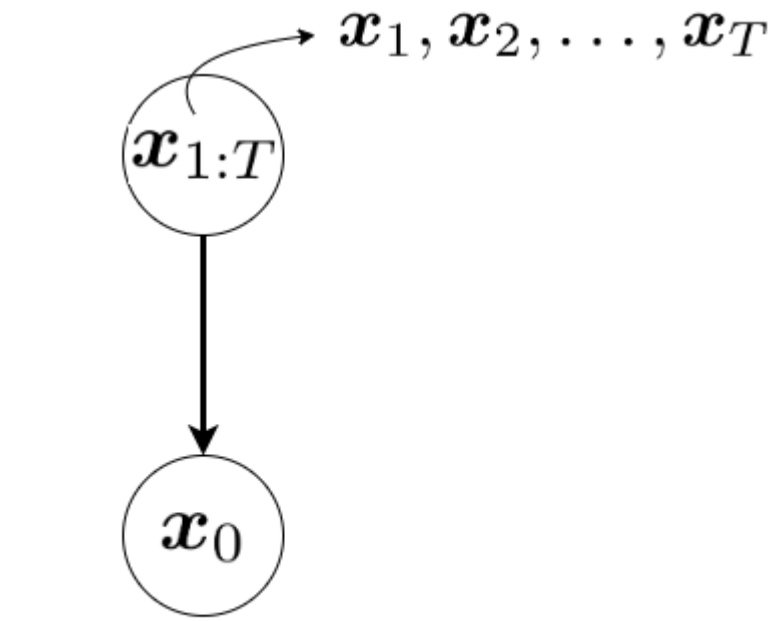
Joint Distribution over Latent States in DDPMs

The joint distribution is expressed as:

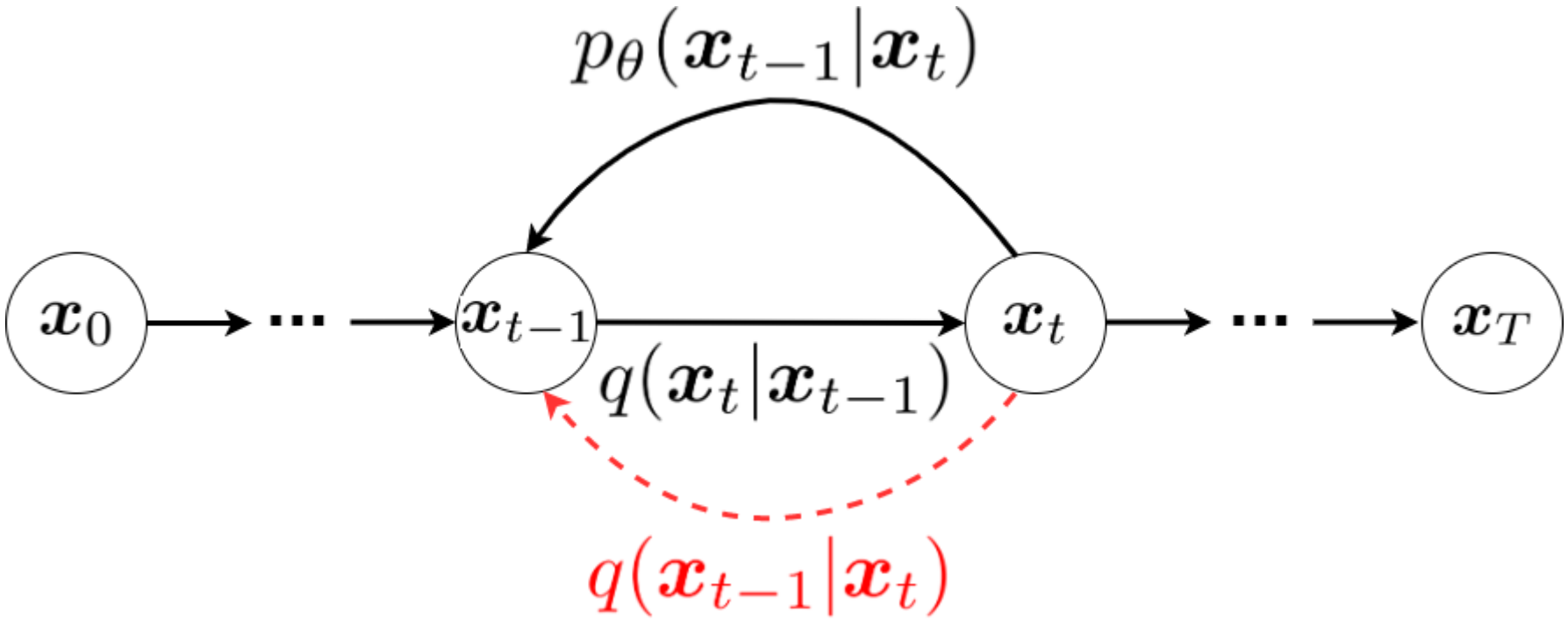
$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$$

Where:

- $x_{1:T}$ represents all latent states from time 1 to T.
- x_0 is the initial state.



Diffusion model


$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$
$\boldsymbol{x}_0$
$\boldsymbol{x}_{t-1}$
$\boldsymbol{x}_t$
$\boldsymbol{x}_T$
$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$
$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$

$$x_0 \longleftarrow x_1 \longleftarrow \cdots \longleftarrow x_T$$

The generator is modeled to reverse the diffusion process.

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}|\mu_\theta(x_t;t), \Sigma_\theta(x_t;t))$$

where the variance $\Sigma_\theta(x_t;t)$ is often set to $\sigma_t^2 I$.

$$x_0 \longleftarrow x_1 \longleftarrow \cdots \longleftarrow x_T$$

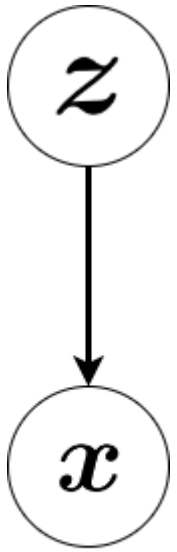
The joint distribution is given by

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$$

with $p(x_T)$ being $\mathcal{N}(0; I)$.

DDPM vs VAE

- **In VAE**
  - Observed variable x
  - Hidden Variable z



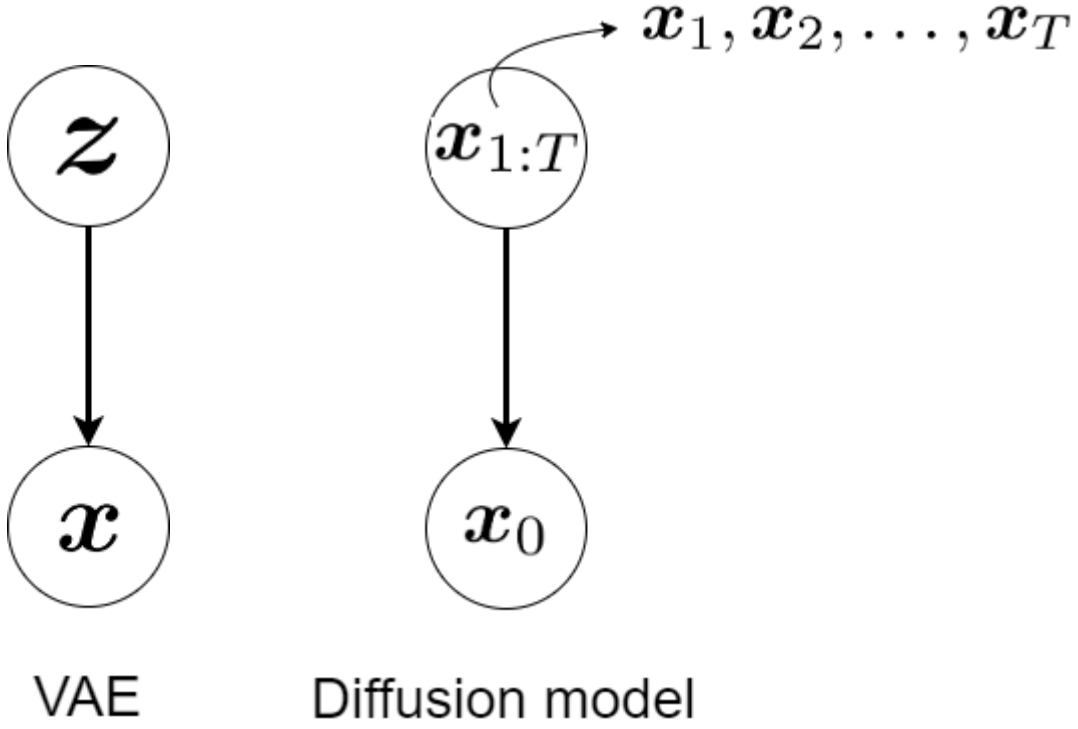
VAE

DDPM vs VAE

- In VAE
  - Observed variable x
  - Hidden Variable z

$$x_0 \longleftarrow x_1 \longleftarrow \cdots \longleftarrow x_T$$

- **In DDPM**
  - Observed variable x_0
  - Hidden variable $x_1, x_2, \ldots x_T$



Objective Functions in VAEs vs. DDPMs

VAEs:
Maximize the likelihood of the data:

$$p_\theta(x) = \int p_\theta(x, z)\, dz$$

Objective Functions in VAEs vs. DDPMs

VAEs:
Maximize the likelihood of the data:

$$p_\theta(x) = \int p_\theta(x, z)\, dz$$

Diffusion Models:

$$p_\theta(x_0) = \int p_\theta(x_0, x_{1:T})\, dx_{1:T}$$

Variational Lower Bounds in VAEs vs. DDPMs

VAEs:

$$p_\theta(x) \geq \mathbb{E}_{q(z|x)}[\log p_\theta(x|z)] - D_{KL}(q(z|x) \,||\, p_\theta(z))$$

Variational Lower Bounds in VAEs vs. DDPMs

VAEs:

$$p_\theta(x) \geq \mathbb{E}_{q(z|x)}[\log p_\theta(x|z)] - D_{KL}(q(z|x) \,||\, p_\theta(z))$$

Diffusion Models:

$$p_\theta(x_0) \geq \mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - D_{KL}(q(x_{1:T}|x_0) \,||\, p_\theta(x_{1:T}))$$

Variational Lower Bound

$$\mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}))$$

Variational Lower Bound

$$\mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}))$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - \int q(x_{1:T}|x_0) \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})} dx_{1:T}$$

Variational Lower Bound

$$\mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}))$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - \int q(x_{1:T}|x_0) \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})} dx_{1:T}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})} \right]$$

Variational Lower Bound

$$\mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}))$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - \int q(x_{1:T}|x_0) \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})} dx_{1:T}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})}\right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log p_\theta(x_0|x_{1:T}) - \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})}\right]$$

Variational Lower Bound

$$\mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}))$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - \int q(x_{1:T}|x_0) \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})} dx_{1:T}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})}\right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log p_\theta(x_0|x_{1:T}) - \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})}\right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log p_\theta(x_0|x_{1:T}) + \log \frac{p_\theta(x_{1:T})}{q(x_{1:T}|x_0)}\right]$$

Variational Lower Bound

$$\mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}))$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - \int q(x_{1:T}|x_0) \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})} dx_{1:T}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})}\right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log p_\theta(x_0|x_{1:T}) - \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})}\right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log p_\theta(x_0|x_{1:T}) + \log \frac{p_\theta(x_{1:T})}{q(x_{1:T}|x_0)}\right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \frac{p_\theta(x_0|x_{1:T})p_\theta(x_{1:T})}{q(x_{1:T}|x_0)}\right] = \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right]$$

$$\mathbb{E}_q \left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right]$$

$$\mathbb{E}_q \left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right]$$

$$= \mathbb{E}_q \left[\log \frac{p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{q(x_{1:T}|x_0)} \right]$$

$$\mathbb{E}_q \left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right]$$

$$= \mathbb{E}_q \left[\log \frac{p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{q(x_{1:T}|x_0)} \right]$$

$$= \mathbb{E}_q[\log p(x_T)] + \sum_{t=1}^{T} \mathbb{E}_q[\log p_\theta(x_{t-1}|x_t)] - \mathbb{E}_q[\log q(x_{1:T}|x_0)]$$

$$\mathbb{E}_q \left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right]$$

$$= \mathbb{E}_q \left[\log \frac{p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{q(x_{1:T}|x_0)} \right]$$

$$= \mathbb{E}_q[\log p(x_T)] + \sum_{t=1}^{T} \mathbb{E}_q[\log p_\theta(x_{t-1}|x_t)] - \mathbb{E}_q[\log q(x_{1:T}|x_0)]$$

$$= \mathbb{E}_q[\log p(x_T)] + \sum_{t=1}^{T} \mathbb{E}_q[\log p_\theta(x_{t-1}|x_t)] - \mathbb{E}_q \left[\log \prod_{t=1}^{T} q(x_t|x_{t-1}) \right]$$

$$\mathbb{E}_q \left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right]$$

$$= \mathbb{E}_q \left[\log \frac{p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{q(x_{1:T}|x_0)} \right]$$

$$= \mathbb{E}_q[\log p(x_T)] + \sum_{t=1}^{T} \mathbb{E}_q[\log p_\theta(x_{t-1}|x_t)] - \mathbb{E}_q[\log q(x_{1:T}|x_0)]$$

$$= \mathbb{E}_q[\log p(x_T)] + \sum_{t=1}^{T} \mathbb{E}_q[\log p_\theta(x_{t-1}|x_t)] - \mathbb{E}_q \left[\log \prod_{t=1}^{T} q(x_t|x_{t-1}) \right]$$

$$= \mathbb{E}_q[\log p(x_T)] + \sum_{t=1}^{T} \mathbb{E}_q[\log p_\theta(x_{t-1}|x_t)] - \sum_{t=1}^{T} \mathbb{E}_q[\log q(x_t|x_{t-1})]$$

$$\mathbb{E}_q \left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right]$$

$$= \mathbb{E}_q \left[\log \frac{p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{q(x_{1:T}|x_0)} \right]$$

$$= \mathbb{E}_q[\log p(x_T)] + \sum_{t=1}^{T} \mathbb{E}_q[\log p_\theta(x_{t-1}|x_t)] - \mathbb{E}_q[\log q(x_{1:T}|x_0)]$$

$$= \mathbb{E}_q[\log p(x_T)] + \sum_{t=1}^{T} \mathbb{E}_q[\log p_\theta(x_{t-1}|x_t)] - \mathbb{E}_q \left[\log \prod_{t=1}^{T} q(x_t|x_{t-1}) \right]$$

$$= \mathbb{E}_q[\log p(x_T)] + \sum_{t=1}^{T} \mathbb{E}_q[\log p_\theta(x_{t-1}|x_t)] - \sum_{t=1}^{T} \mathbb{E}_q[\log q(x_t|x_{t-1})]$$

$$= \mathbb{E}_q[\log p(x_T)] + \sum_{t=1}^{T} \mathbb{E}_q \left[\log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right]$$

Decomposing the ELBO

- $\mathbb{E}_q[\log p(x_T)]$: Validates the model's calibration to the true noise at the last diffusion step, ensuring the reverse process starts accurately.

- $\sum_{t=1}^{T} \mathbb{E}_q \left[\log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right]$: Acts as a regularizer by comparing the model's backward predictions to the known forward diffusion, guiding the model to correct any discrepancies.

$$\mathcal{L}(\boldsymbol{x}_0) = \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log p(\boldsymbol{x}_T) + \sum_{t=1}^{T} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right]$$

$$\mathcal{L}(\boldsymbol{x}_0) = \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log p(\boldsymbol{x}_T) + \sum_{t=1}^{T} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right]$$

By the Markov property:

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{x}_0),$$

$$\mathcal{L}(\boldsymbol{x}_0) = \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log p(\boldsymbol{x}_T) + \sum_{t=1}^{T} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right]$$

By the Markov property:

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{x}_0),$$

and by Bayes' rule:

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{x}_0) = \frac{q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_0)}.$$

$$\mathcal{L}(\boldsymbol{x}_0) = \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log p(\boldsymbol{x}_T) + \sum_{t=1}^{T} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right]$$

By the Markov property:

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{x}_0),$$

and by Bayes' rule:

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{x}_0) = \frac{q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_0)}.$$

Plugging this equation into the ELBO, we get

$$\ell(\boldsymbol{x}_0) = \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log p(\boldsymbol{x}_T) + \sum_{t=2}^{T} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0)} \right.$$

$$\left. + \underbrace{\sum_{t=2}^{T} \log \frac{q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_0)}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)}}_{*} + \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \right]$$

$$\underbrace{\sum_{t=2}^{T} \log \frac{q\left(x_{t-1} \mid x_0\right)}{q\left(x_t \mid x_0\right)}}_{*}$$

The term marked * is a telescoping sum, and can be simplified as follows:

$$\begin{aligned} * &= \log q\left(x_{T-1} \mid x_0\right) + \cdots + \log q\left(x_2 \mid x_0\right) + \log q\left(x_1 \mid x_0\right) \\ &\quad - \log q\left(x_T \mid x_0\right) - \log q\left(x_{T-1} \mid x_0\right) - \cdots - \log q\left(x_2 \mid x_0\right) \\ &= -\log q\left(x_T \mid x_0\right) + \log q\left(x_1 \mid x_0\right) \end{aligned}$$

$$\mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}[\log p\left(\boldsymbol{x}_T\right) \quad + \quad \sum_{t=2}^{T} \log \tfrac{p_{\boldsymbol{\theta}}(x_{t-1}|x_t)}{q(x_{t-1}|x_t,x_0)} \quad +$$

$$\underbrace{\sum_{t=2}^{T} \log \frac{q\left(x_{t-1} \mid x_0\right)}{q\left(x_t \mid x_0\right)}}_{*} + \log \tfrac{p_{\boldsymbol{\theta}}(x_0|x_1)}{q(x_1|x_0)}]$$

$$* = -\log q\left(x_T \mid x_0\right) + \log q\left(x_1 \mid x_0\right)$$

Hence the negative ELBO (variational upper bound) becomes

$$= -\mathbb{E}_q \left[\log \frac{p\left(\boldsymbol{x}_T\right)}{q\left(\boldsymbol{x}_T \mid \boldsymbol{x}_0\right)} + \sum_{t=2}^{T} \log \frac{p_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)}{q\left(x_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0\right)} + \log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}_0 \mid \boldsymbol{x}_1\right)\right]$$

$$= \underbrace{D_{\mathrm{KL}}\left(q\left(\boldsymbol{x}_T \mid \boldsymbol{x}_0\right) \| p\left(\boldsymbol{x}_T\right)\right)}_{L_T(\boldsymbol{x}_0)}$$

$$+ \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \underbrace{D_{\mathbb{KL}}\left(q\left(x_{t-1} \mid x_t, x_0\right) \| p_{\boldsymbol{\theta}}\left(x_{t-1} \mid x_t\right)\right)}_{L_{t-1}(\boldsymbol{x}_0)}$$

$$- \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \log p_{\boldsymbol{\theta}}\left(x_0 \mid x_1\right)}_{L_0(\boldsymbol{x}_0)}$$

Revers Process

We need to learn a neural network to approximate the conditioned probability distributions in the reverse diffusion process,

$$p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)=\mathcal{N}\left(\mathbf{x}_{t-1} ; \boldsymbol{\mu}_\theta\left(\mathbf{x}_t, t\right), \boldsymbol{\Sigma}_\theta\left(\mathbf{x}_t, t\right)\right).$$

$q(\mathbf{x}_t \mid \mathbf{x}_0)$

Recall that the forward process is defined by the equation:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t \mid \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

We can sample $\mathbf{x}_t$ at any arbitrary time step t in a closed form.
Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$:

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

Mean of $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)$

The reverse conditional probability is tractable when conditioned on $\mathbf{x}_0$:

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

Using Bayes' rule, we have:

$$\begin{aligned}
q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) &= q(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{x}_0) \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_0)}{q(\mathbf{x}_t \mid \mathbf{x}_0)} \\
&\propto \exp\left(-\frac{1}{2}\left(\frac{\left(\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}_{t-1}\right)^2}{\beta_t} + \frac{\left(\mathbf{x}_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0\right)^2}{1-\bar{\alpha}_{t-1}} - \frac{\left(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0\right)^2}{1-\bar{\alpha}_t}\right)\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{\mathbf{x}_t^2 - 2\sqrt{\alpha_t}\mathbf{x}_t\mathbf{x}_{t-1} + \alpha_t \mathbf{x}_{t-1}^2}{\beta_t} + \frac{\mathbf{x}_{t-1}^2 - 2\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0\mathbf{x}_{t-1} + \bar{\alpha}_{t-1}\mathbf{x}_0^2}{1-\bar{\alpha}_{t-1}} - \frac{\left(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0\right)^2}{1-\bar{\alpha}_t}\right)\right) \\
&= \exp\left(-\frac{1}{2}\left(\left(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}\right)\mathbf{x}_{t-1}^2 - \left(\frac{2\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t + \frac{2\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\mathbf{x}_0\right)\mathbf{x}_{t-1} + C(\mathbf{x}_t, \mathbf{x}_0)\right)\right)
\end{aligned}$$

where $C(\mathbf{x}_t, \mathbf{x}_0)$ is some function not involving $\mathbf{x}_{t-1}$.

Mean of $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)$

We can express the mean and variance in the following manner: (recall that $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i$) :

$$\tilde{\beta}_t = 1/\left(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}\right) = 1/\left(\frac{\alpha_t - \bar{\alpha}_t + \beta_t}{\beta_t\left(1-\bar{\alpha}_{t-1}\right)}\right) = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t$$

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

Recall $\mathbf{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t\right)$

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}_t &= \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t\right) \\
&= \frac{1}{\sqrt{\alpha_t}}\left(\mathrm{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t\right)
\end{aligned}$$

KL for two Guassian

$$D_{KL}\left(q\left(x_{t-1} \mid x_t, x_0\right) \| p_{\boldsymbol{\theta}}\left(x_{t-1} \mid x_t\right)\right)$$

$$q(x) = \mathcal{N}(x; \mu_q, \Sigma_q) \quad \text{and} \quad p(x) = \mathcal{N}(x; \mu_p, \Sigma_p),$$

$$D_{KL}(q\|p) = \frac{1}{2}\left(\operatorname{tr}(\Sigma_p^{-1}\Sigma_q) + (\mu_p - \mu_q)^T\Sigma_p^{-1}(\mu_p - \mu_q) - k + \ln\left(\frac{\det\Sigma_p}{\det\Sigma_q}\right)\right)$$

for q

$$\tilde{\boldsymbol{\mu}}_t\left(\boldsymbol{x}_t, \boldsymbol{x}_0\right) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_+ - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right)$$

for p

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_t, t\right) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_t, t\right)\right)$$

Loss fnction

$$
\begin{aligned}
L_t &= E_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{1}{2\left\|\boldsymbol{\Sigma}_\theta\left(\mathbf{x}_t,t\right)\right\|_2^2}\left\|\tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t,\mathbf{x}_0\right)-\boldsymbol{\mu}_\theta\left(\mathbf{x}_t,t\right)\right\|^2\right] \\
&= E_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{1}{2\left\|\boldsymbol{\Sigma}_\theta\right\|_2^2}\left\|\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_t\right)-\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t,t\right)\right)\right\|^2\right] \\
&= E_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{\left(1-\alpha_t\right)^2}{2\alpha_t\left(1-\bar{\alpha}_t\right)\left\|\boldsymbol{\Sigma}_\theta\right\|_2^2}\left\|\boldsymbol{\epsilon}_t-\boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t,t\right)\right\|^2\right] \\
&= E_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{\left(1-\alpha_t\right)^2}{2\alpha_t\left(1-\bar{\alpha}_t\right)\left\|\boldsymbol{\Sigma}_\theta\right\|_2^2}\left\|\boldsymbol{\epsilon}_t-\boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t,t\right)\right\|^2\right]
\end{aligned}
$$

Loss function

- Ho et al. (2020) discovered empirically that the diffusion model produces higher-quality images when using a simplified objective, omiting the weighting term

$$
\begin{aligned}
L_{\text{simple}} &= E_{t\sim[1,T],\mathbf{x}_0,\epsilon_t}\left[\left\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t\right)\right\|^2\right] \\
&= E_{t\sim[1,T],\mathbf{x}_0,\epsilon_t}\left[\left\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t, t\right)\right\|^2\right]
\end{aligned}
$$

Training a DDPM model with L_{simple}.

while *not converged* **do**

- $\boldsymbol{x}_0 \sim q_0(\boldsymbol{x}_0)$
- $t \sim \text{Unif}(\{1, \ldots, T\})$
- $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- Take gradient descent step on $\nabla_{\boldsymbol{\theta}} ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}} \left(\sqrt{\overline{\alpha}_t} \boldsymbol{x}_0 + \sqrt{1 - \overline{\alpha}_t} \boldsymbol{\epsilon}, t \right) ||^2$

Training a DDPM model with L_{simple}.

while *not converged* **do**
- $\boldsymbol{x}_0 \sim q_0(\boldsymbol{x}_0)$
- $t \sim \text{Unif}(\{1, \dots, T\})$
- $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- Take gradient descent step on $\nabla_{\boldsymbol{\theta}} ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}} \left(\sqrt{\overline{\alpha}_t} \boldsymbol{x}_0 + \sqrt{1 - \overline{\alpha}_t} \boldsymbol{\epsilon}, t \right) ||^2$

Sampling from a DDPM model.

$\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

foreach $t = T, \dots, 1$ **do**
- $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- $\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}} \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) \right) + \sigma_t \boldsymbol{\epsilon}_t$

Return $\boldsymbol{x}_0$