Deep Generative Models: Diffusion Models

Fall Semester 2026

René Vidal
Director of the Center for Innovation in Data Engineering and Science (IDEAS)
Rachleff University Professor, University of Pennsylvania
Amazon Scholar & Chief Scientist at NORCE

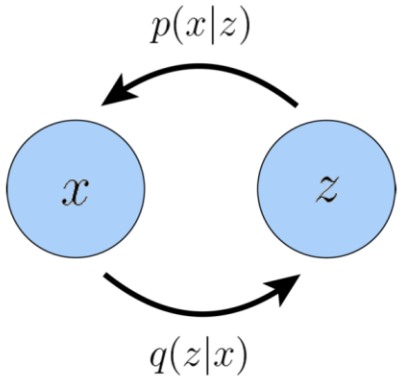


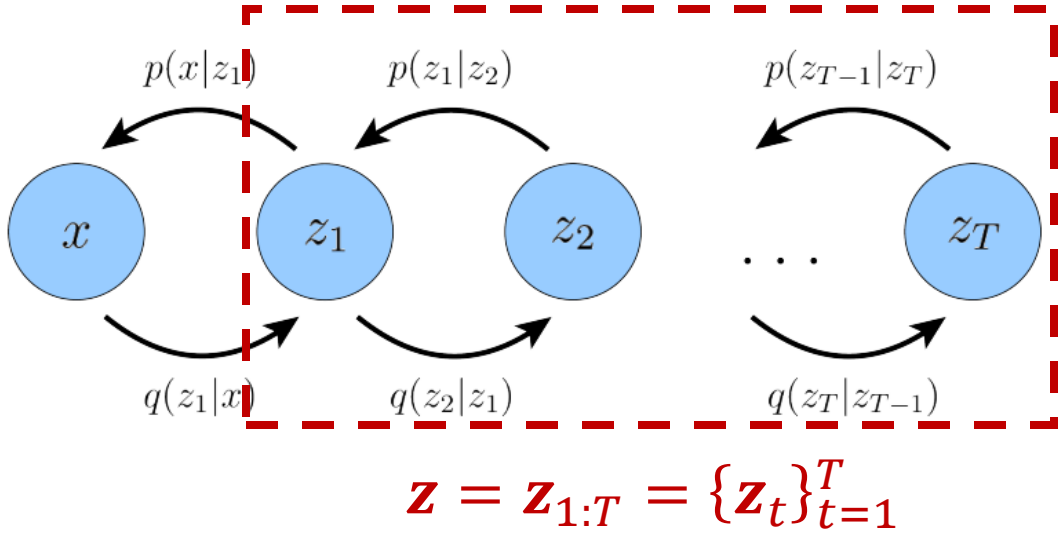
Outline

- Markov Hierarchical Variational Auto Encoders (MHVAEs)
  - Autoregressive Encoder and Autoregressive Decoder of an MHVAE
  - Derivation of the ELBO of an MHVAE
- **Diffusion Models: MHVAEs with a Linear Gaussian Autoregressive Latent Space**
  - Forward Diffusion Process
  - Reverse Diffusion Process
  - ELBO for Diffusion Models as a particular case of the ELBO for MHVAEs
- Implementation Details: UNet architecture, Training and Sampling Strategies
- Application of Diffusion Models
  - Text-Conditioned Diffusion Model: Stable Diffusion
  - Multimodal Control for Consistent Synthesis: ControlNet
  - Image Editing: DDIM, P2P

Summary of Last Lecture

- A Latent Diffusion Model is as a Markovian Hierarchical Variational Autoencoder (MHVAE) with T hierarchical latents $\boldsymbol{z} = \boldsymbol{z}_{1:T} = \{\boldsymbol{z}_t\}_{t=1}^{T}$ modeled by a Markov chain where each latent $\boldsymbol{z}_t$ is generated only from the previous latent $\boldsymbol{z}_{t+1}$.





$\boldsymbol{z} = \boldsymbol{z}_{1:T} = \{\boldsymbol{z}_t\}_{t=1}^{T}$

- The ELBO of a Diffusion Model can be written as

$$\mathcal{L}_{\text{ELBO}}(\theta, \phi)$$
$$= \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{z}_1|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}_1)]}_{\text{reconstruction term}} - \underbrace{D_{\text{KL}}\left(q_\phi(\boldsymbol{z}_T \mid \boldsymbol{x}) \,||\, p_\theta(\boldsymbol{z}_T)\right)}_{\text{prior matching term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{z}_t|\boldsymbol{x})}\left[D_{\text{KL}}\left(q_\phi(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t, \boldsymbol{x}) \,||\, p_\theta(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t)\right)\right]}_{\text{score matching term}}$$

Diffusion Model as MHVAEs with Gaussian Latents

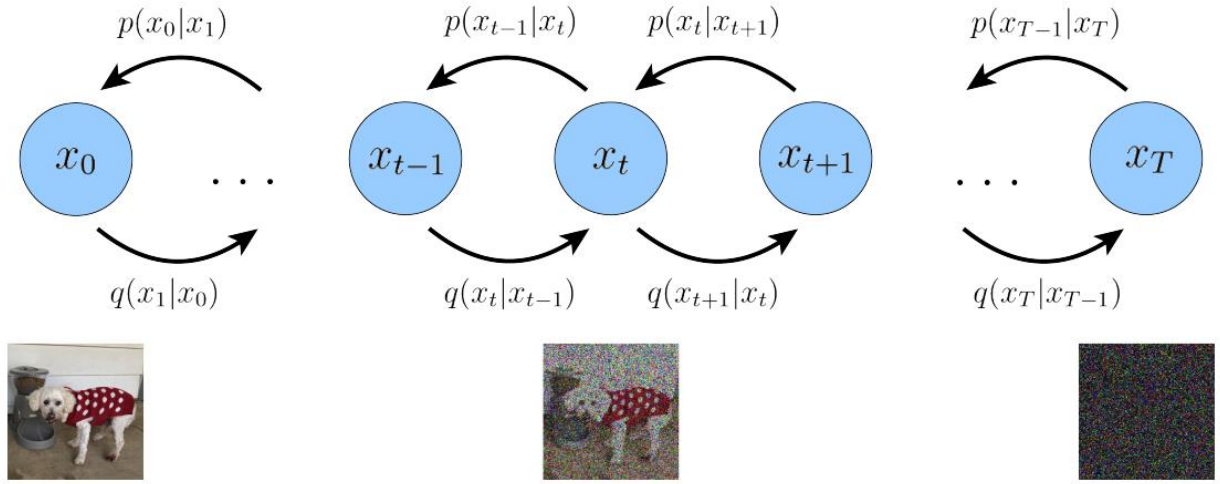
- **A Diffusion Model is an MHVAE** where the latent variables $\boldsymbol{x}_{1:T}$ have the same dimension as the data $\boldsymbol{x}_0$, and the encoder $q_\phi(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0) = \prod_{t=1}^{T} q_\phi(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$ is not learned, but it is pre-specified as a linear Gaussian model

$$q_\phi(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\boldsymbol{I})$$

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

  where $\boldsymbol{\epsilon}_t$ are i.i.d. and independent of $\boldsymbol{x}_{t-1}$.

- The parameter α_t is chosen such that $\boldsymbol{x}_T \sim \mathcal{N}(\boldsymbol{x}_T; \mathbf{0}, \boldsymbol{I})$ is a standard Gaussian



The Forward Process of a Diffusion Model

- Consider the formulation of a single noising step:

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

- Since $\boldsymbol{x}_1 \mid \boldsymbol{x}_0$ and $\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}$ are Gaussian, so is $\boldsymbol{x}_t \mid \boldsymbol{x}_0 \sim \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_q(\boldsymbol{x}_0), \boldsymbol{\Sigma}_q(\boldsymbol{x}_0))$.
- We can compute $\boldsymbol{\mu}_q(\boldsymbol{x}_0) = \mathbb{E}[\,\boldsymbol{x}_t \mid \boldsymbol{x}_0\,]$, recursively as follows:

$$\begin{aligned}
\mathbb{E}[\,\boldsymbol{x}_t \mid \boldsymbol{x}_0\,] &= \mathbb{E}\big[\,\sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t \mid \boldsymbol{x}_0\,\big] \\
&= \sqrt{\alpha_t}\,\mathbb{E}[\,\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_0\,] + \sqrt{1-\alpha_t}\mathbb{E}[\boldsymbol{\epsilon}_t] \\
&= \sqrt{\alpha_t}\mathbb{E}[\,\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_0\,] \\
&= \sqrt{\alpha_t}\sqrt{\alpha_{t-1}}\,\mathbb{E}[\,\boldsymbol{x}_{t-2} \mid \boldsymbol{x}_0\,] \\
&= \sqrt{\alpha_t}\sqrt{\alpha_{t-1}}\cdots\sqrt{\alpha_1}\,\boldsymbol{x}_0 \\
&= \sqrt{\overline{\alpha_t}}\boldsymbol{x}_0
\end{aligned}$$

$\mathbb{E}[\boldsymbol{\epsilon}_t] = \mathbf{0}$

$$\overline{\alpha_t} = \prod_{i=1}^{t} \alpha_i$$

The Forward Process of a Diffusion Model

- We can compute $\boldsymbol{\Sigma}_q(\boldsymbol{x}_0) = \mathrm{Var}[\,\boldsymbol{x}_t \mid \boldsymbol{x}_0\,]$, recursively as follows:

$$\begin{aligned}\mathrm{Var}(\,\boldsymbol{x}_t \mid \boldsymbol{x}_0\,) &= \mathrm{Var}\big(\sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t \mid \boldsymbol{x}_0\big) \\ &= \alpha_t \mathrm{Var}(\,\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_0\,) + (1-\alpha_t)\,\mathrm{Var}(\boldsymbol{\epsilon}_t) \\ &= \alpha_t \mathrm{Var}(\,\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_0\,) + (1-\alpha_t)\,\boldsymbol{I}\end{aligned}$$

- That is:

$$\begin{aligned}\mathrm{Var}(\,\boldsymbol{x}_t \mid \boldsymbol{x}_0\,) &= \alpha_t\,[\alpha_{t-1}\mathrm{Var}(\,\boldsymbol{x}_{t-2} \mid \boldsymbol{x}_0\,) + (1-\alpha_{t-1})\,\boldsymbol{I}] + (1-\alpha_t)\,\boldsymbol{I} \\ &= \alpha_t\alpha_{t-1}\mathrm{Var}(\,\boldsymbol{x}_{t-2} \mid \boldsymbol{x}_0\,) + (1-\alpha_t\alpha_{t-1})\boldsymbol{I} \\ &= \cdots \\ &= \alpha_t\alpha_{t-1}\ldots\alpha_1\,\mathrm{Var}(\,\boldsymbol{x}_0 \mid \boldsymbol{x}_0\,) + \left(1 - \prod_{i=1}^{t}\alpha_i\right)\boldsymbol{I} \\ &= \left(1 - \textstyle\prod_{i=1}^{t}\alpha_i\right)\boldsymbol{I} = (1-\overline{\alpha_t})\boldsymbol{I}\end{aligned}$$

$\mathrm{Var}(\,\boldsymbol{x}_0 \mid \boldsymbol{x}_0\,) \to \mathbf{0}$

$$\overline{\alpha_t} = \prod_{i=1}^{t}\alpha_i$$

The Forward Process of a Diffusion Model

- Therefore, given $\boldsymbol{x}_0$, we can sample $\boldsymbol{x}_t$ directly without having to generate all $\boldsymbol{x}_t$'s:

$$\boldsymbol{x}_t = \sqrt{\overline{\alpha_t}}\boldsymbol{x}_0 + \sqrt{1-\overline{\alpha_t}}\,\overline{\boldsymbol{\epsilon}}_t, \qquad \overline{\boldsymbol{\epsilon}}_t \sim \mathcal{N}(\overline{\boldsymbol{\epsilon}}_t; \mathbf{0}, \boldsymbol{I})$$

- We can compute $\boldsymbol{\mu}_q(\boldsymbol{x}_0) = \mathbb{E}[\,\boldsymbol{x}_t \mid \boldsymbol{x}_0\,]$, recursively as follows:

$$\begin{aligned}\boldsymbol{x}_t &= \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\,\boldsymbol{\epsilon}_t = \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\,\boldsymbol{\epsilon}_{t-1}\right) + \sqrt{1-\alpha_t}\,\boldsymbol{\epsilon}_t \\ &= \sqrt{\alpha_t}\sqrt{\alpha_{t-1}}\cdots\sqrt{\alpha_1}\,\boldsymbol{x}_0 \\ &= \sqrt{\overline{\alpha_t}}\boldsymbol{x}_0\end{aligned}$$

$\boldsymbol{\epsilon}_{t-1}$ → $\mathbf{0}$

$$\overline{\alpha_t} = \prod_{i=1}^{t} \alpha_i$$

The Forward Process of Diffusion Model

- We have shown that $\boldsymbol{x}_t$ is a linear Gaussian transformation of $\boldsymbol{x}_0$ with scheduled randomness (controlled by $\overline{\alpha_t}$) drawn from a standard normal distribution, i.e.,

$$\boldsymbol{x}_t \mid \boldsymbol{x}_0 \sim \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\overline{\alpha_t}}\boldsymbol{x}_0, (1-\overline{\alpha_t})\boldsymbol{I}\right)$$

- Therefore, given $\boldsymbol{x}_0$, we can sample $\boldsymbol{x}_t$ directly without having to generate all $\boldsymbol{x}_t$'s:

$$\boldsymbol{x}_t = \sqrt{\overline{\alpha_t}}\boldsymbol{x}_0 + \sqrt{1-\overline{\alpha_t}}\,\overline{\boldsymbol{\epsilon}_t}, \qquad \overline{\boldsymbol{\epsilon}_t} \sim \mathcal{N}(\overline{\boldsymbol{\epsilon}_t}; \mathbf{0}, \boldsymbol{I})$$

- Moreover, we can also generate $\boldsymbol{x}_0$ from $\boldsymbol{x}_t$ as

$$\boldsymbol{x}_0 = (\boldsymbol{x}_t - \sqrt{1-\overline{\alpha_t}}\,\overline{\boldsymbol{\epsilon}_t})/\sqrt{\overline{\alpha_t}}, \qquad \overline{\boldsymbol{\epsilon}_t} \sim \mathcal{N}(\overline{\boldsymbol{\epsilon}_t}; \mathbf{0}, \boldsymbol{I})$$

- This suggests we can reverse the noising process. However, exact reversal requires knowing the exact $\overline{\boldsymbol{\epsilon}_t}$. The reverse diffusion process is designed to predict the noise $\overline{\boldsymbol{\epsilon}_t}$ that needs to be subtracted from $\boldsymbol{x}_t$ to generate $\boldsymbol{x}_0$.

The Reverse Diffusion Process

- We have designed a forward diffusion process $q_\phi(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$ that
  - At each step adds Gaussian noise to the input until it becomes pure noise
  - Allows us to sample $\boldsymbol{x}_t \mid \boldsymbol{x}_0$ without having to compute $\boldsymbol{x}_t$ recursively
  - Allows us to sample $\boldsymbol{x}_0 \mid \boldsymbol{x}_t$ without having to compute $\boldsymbol{x}_0$ recursively

- We now need to design a reverse diffusion process $p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ that makes the calculation of the ELBO easy. We do this by
  - Understanding how $q_\phi(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0)$ depends on $\boldsymbol{x}_t, \boldsymbol{x}_0$
  - Designing $p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ so that it matches that dependence

- Recall the ELBO is given by:

$$= \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{x}_T \mid \boldsymbol{x}_0) \,||\, p_\theta(\boldsymbol{x}_T)\right)}_{\text{prior matching term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \,||\, p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)\right)\right]}_{\text{score matching term}}$$

ELBO for Diffusion Model: Score Matching Term

- To compute the third term, we need

$$q_\phi(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{q_\phi(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{x}_0) q_\phi(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_0)}{q_\phi(\boldsymbol{x}_t \mid \boldsymbol{x}_0)} = \frac{\mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\boldsymbol{I}\right)\mathcal{N}\left(\boldsymbol{x}_{t-1}; \sqrt{\overline{\alpha_{t-1}}}\boldsymbol{x}_0, (1-\overline{\alpha_{t-1}})\boldsymbol{I}\right)}{\mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\overline{\alpha_t}}\boldsymbol{x}_0, (1-\overline{\alpha_t})\boldsymbol{I}\right)}$$

$$\propto \exp\left(-\frac{||\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_{t-1}||^2}{2(1-\alpha_t)}\right) \cdot \exp\left(-\frac{||\boldsymbol{x}_{t-1} - \sqrt{\overline{\alpha_{t-1}}}\boldsymbol{x}_0||^2}{2(1-\overline{\alpha_{t-1}})}\right) \propto \exp\left(-\frac{\alpha_t ||\boldsymbol{x}_{t-1} - \frac{\boldsymbol{x}_t}{\sqrt{\alpha_t}}||^2}{2(1-\alpha_t)}\right) \cdot \exp\left(-\frac{||\boldsymbol{x}_{t-1} - \sqrt{\overline{\alpha_{t-1}}}\boldsymbol{x}_0||^2}{2(1-\overline{\alpha_{t-1}})}\right)$$

- Applying the product rule, we get $q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_q(\boldsymbol{x}_t, \boldsymbol{x}_0), \boldsymbol{\Sigma}_q)$, where

$$\boldsymbol{\Sigma}_q := \mathrm{Cov}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) = \left(\frac{\alpha_t}{1-\alpha_t}\boldsymbol{I} + \frac{1}{1-\overline{\alpha_{t-1}}}\boldsymbol{I}\right)^{-1} = \frac{(1-\alpha_t)(1-\overline{\alpha_{t-1}})}{1-\overline{\alpha_t}}\boldsymbol{I}$$

$$\boldsymbol{\mu}_q(\boldsymbol{x}_t, \boldsymbol{x}_0) := \mathbb{E}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) = \boldsymbol{\Sigma}_q\left(\frac{\alpha_t}{1-\alpha_t}\boldsymbol{I}\,\frac{\boldsymbol{x}_t}{\sqrt{\alpha_t}} + \frac{1}{1-\overline{\alpha_{t-1}}}\boldsymbol{I}\sqrt{\overline{\alpha_{t-1}}}\boldsymbol{x}_0\right)$$

$$= \frac{(1-\alpha_t)(1-\overline{\alpha_{t-1}})}{1-\overline{\alpha_t}}\boldsymbol{I}\left(\frac{\sqrt{\alpha_t}}{1-\alpha_t}\boldsymbol{x}_t + \frac{\sqrt{\overline{\alpha_{t-1}}}}{1-\overline{\alpha_{t-1}}}\boldsymbol{x}_0\right) = \frac{\sqrt{\alpha_t}(1-\overline{\alpha_{t-1}})\boldsymbol{x}_t + \sqrt{\overline{\alpha_{t-1}}}(1-\alpha_t)\boldsymbol{x}_0}{1-\overline{\alpha_t}}$$

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\boldsymbol{I})$$
$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\overline{\alpha_t}}\boldsymbol{x}_0, (1-\overline{\alpha_t})\boldsymbol{I})$$

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\,\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \propto \mathcal{N}(\boldsymbol{x}; \overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\Sigma}})$$
$$\overline{\boldsymbol{\mu}} = \overline{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2), \overline{\boldsymbol{\Sigma}} = (\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})^{-1}$$

$$\boldsymbol{\mu}_1 = \frac{\boldsymbol{x}_t}{\sqrt{\alpha_t}}, \boldsymbol{\Sigma}_1 = \frac{1-\alpha_t}{\alpha_t}\boldsymbol{I},$$
$$\boldsymbol{\mu}_2 = \sqrt{\overline{\alpha_{t-1}}}\boldsymbol{x}_0, \quad \boldsymbol{\Sigma}_2 = (1-\overline{\alpha_{t-1}})\boldsymbol{I}$$

ELBO for Diffusion Model: Decoder Matches Encoder

- Recall KL divergence for Gaussians

$$\mathrm{D}_{\mathrm{KL}}\left(\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q)\,||\,\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_\theta,\boldsymbol{\Sigma}_\theta)\right)=\frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_\theta|}{|\boldsymbol{\Sigma}_q|}-d+\mathrm{tr}\left(\boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{\Sigma}_q\right)+\left(\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_q\right)^{\mathrm{T}}\boldsymbol{\Sigma}_\theta^{-1}\left(\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_q\right)\right]$$

- Choose mean of $p_\theta(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t)$ to match the form the of mean of $q_\phi(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t,\boldsymbol{x}_0)$

$$\boldsymbol{\mu}_q(\boldsymbol{x}_t,\boldsymbol{x}_0)=\frac{\sqrt{\alpha_t}(1-\overline{\alpha_{t-1}})\boldsymbol{x}_t+\sqrt{\overline{\alpha_{t-1}}}(1-\alpha_t)\boldsymbol{x}_0}{1-\overline{\alpha_t}}\quad\Rightarrow\quad\boldsymbol{\mu}_\theta(\boldsymbol{x}_t,t)=\frac{\sqrt{\alpha_t}(1-\overline{\alpha_{t-1}})\boldsymbol{x}_t+\sqrt{\overline{\alpha_{t-1}}}(1-\alpha_t)\widehat{\boldsymbol{x}_\theta}(\boldsymbol{x}_t,t)}{1-\overline{\alpha_t}}$$

- Choose variance of $p_\theta(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t)$ to match exactly the variance of $q(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t,\boldsymbol{x}_0)$

$$\boldsymbol{\Sigma}_q=\frac{(1-\alpha_t)(1-\overline{\alpha_{t-1}})}{1-\overline{\alpha_t}}\boldsymbol{I}\quad\Rightarrow\quad\boldsymbol{\Sigma}_\theta=\frac{(1-\alpha_t)(1-\overline{\alpha_{t-1}})}{1-\overline{\alpha_t}}\boldsymbol{I}$$

$$\sigma_q^2(t)=\frac{(1-\alpha_t)(1-\overline{\alpha_{t-1}})}{1-\overline{\alpha_t}}$$

- The score matching term reduces to:

$$D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t,\boldsymbol{x}_0)\,||\,p_\theta(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t)\right)=D_{\mathrm{KL}}\left(\mathcal{N}(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_q(\boldsymbol{x}_t,\boldsymbol{x}_0),\boldsymbol{\Sigma}_q)||\,\mathcal{N}(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_\theta(\boldsymbol{x}_t,t),\boldsymbol{\Sigma}_q)\right)$$

$$=\frac{1}{2\sigma_q^2(t)}\left[||\,\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_q\,||_2^2\right]=\frac{1-\overline{\alpha_t}}{2(1-\alpha_t)(1-\overline{\alpha_{t-1}})}\frac{\overline{\alpha_{t-1}}(1-\alpha_t)^2}{(1-\overline{\alpha_t})^2}\left[||\widehat{\boldsymbol{x}_\theta}(\boldsymbol{x}_t,t)-\boldsymbol{x}_0||_2^2\right]=\frac{\overline{\alpha_{t-1}}(1-\alpha_t)}{2(1-\overline{\alpha_{t-1}})(1-\overline{\alpha_t})}\left[||\widehat{\boldsymbol{x}_\theta}(\boldsymbol{x}_t,t)-\boldsymbol{x}_0||_2^2\right]$$

Merging Reconstruction and Score Matching Terms

- Can we evaluate the score matching term at $t = 1$?

$$D_{\mathrm{KL}}\left(q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \,||\, p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)\right)$$

- We obtain

$$\begin{aligned} D_{\mathrm{KL}}\left(q(\boldsymbol{x}_0 \mid \boldsymbol{x}_1, \boldsymbol{x}_0) \,||\, p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)\right) &= \int q(\boldsymbol{x}_0 \mid \boldsymbol{x}_1, \boldsymbol{x}_0) \log\left(\frac{q(\boldsymbol{x}_0|\boldsymbol{x}_1, \boldsymbol{x}_0)}{p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)}\right) \\ &= constant - \int q(\boldsymbol{x}_0 \mid \boldsymbol{x}_1, \boldsymbol{x}_0) \log\left(p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)\right) \\ &= constant - \log\left(p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)\right) \end{aligned}$$

- This is because
  - q is constant for optimization purposes
  - $q(\boldsymbol{x}_0 \mid \boldsymbol{x}_1, \boldsymbol{x}_0) = \boldsymbol{\delta}(\boldsymbol{x} - \boldsymbol{x_0})$
- Therefore

$$\mathrm{argmax}_\theta \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)]}_{\text{reconstruction term}} = \mathrm{argmax}_\theta \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[-D_{\mathrm{KL}}\left(q(\boldsymbol{x}_0 \mid \boldsymbol{x}_1, \boldsymbol{x}_0) \,||\, p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)\right)\right]}_{\text{score matching term for t=1}}$$

How to train a Diffusion Model?

$$\sigma_q^2(t) = \begin{cases} \sigma_{\text{rec}}^2, & t = 1 \\ \frac{(1-\alpha_t)(1-\overline{\alpha_{t-1}})}{1-\overline{\alpha_t}}, & t \geq 2 \end{cases}$$

- Recall the ELBO is given by:

$$= \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)]}_{\text{reconstruction term}} - \underbrace{D_{\text{KL}}\left(q_\phi(\boldsymbol{x}_T \mid \boldsymbol{x}_0) \,||\, p_\theta(\boldsymbol{x}_T)\right)}_{\text{prior matching term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\text{KL}}\left(q_\phi(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \,||\, p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)\right)\right]}_{\text{score matching term}}$$

- Since q has no parameters, ELBO reduces to the reconstruction + score matching terms, which we combine to train a network $\widehat{\boldsymbol{x}_\theta}(\boldsymbol{x}_t, t)$ to predict the data $\boldsymbol{x}_0$

$$\underset{\theta}{\text{argmax}}\, \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)]}_{\text{reconstruction term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\frac{\overline{\alpha_{t-1}}(1-\alpha_t)}{2(1-\overline{\alpha_{t-1}})(1-\overline{\alpha_t})}\left[||\widehat{\boldsymbol{x}_\theta}(\boldsymbol{x}_t, t) - \boldsymbol{x}_0||_2^2\right]\right]}_{\text{score matching term}}$$

$$= \underset{\theta}{\text{argmax}} - \sum_{t=1}^{T} \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\text{KL}}\left(q_\phi(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \,||\, p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)\right)\right]}_{\text{reconstruction term} + \text{score matching term}} = -\sum_{t=1}^{T} \mathbb{E}_{q_\phi(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\frac{\overline{\alpha_{t-1}}(1-\alpha_t)^2}{2\sigma_q^2(t)(1-\overline{\alpha}_t)^2}||\widehat{\boldsymbol{x}_\theta}(\boldsymbol{x}_t, t) - \boldsymbol{x}_0||_2^2\right]$$

  - This requires generating $\boldsymbol{x}_t$ given $\boldsymbol{x}_0$ (easy), evaluating $\widehat{\boldsymbol{x}_\theta}(\boldsymbol{x}_t, t)$ and its gradients (easy).
- However, predicting $\boldsymbol{x}_0$ from a highly noisy $\boldsymbol{x}_t$ in one step is complex, as the signal is dominated by significant noise for large t.

Reparameterization as an Alternative Form for ELBO

- Plugging $\boldsymbol{x}_0 = \frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\,\bar{\boldsymbol{\epsilon}}_t}{\sqrt{\bar{\alpha}_t}}$ into the mean $\boldsymbol{\mu}_q(\boldsymbol{x}_t, \boldsymbol{x}_0)$, we obtain:

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

- Choosing the mean and variance of p_θ to match the mean and variance of q_ϕ:

$$\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t, t), \qquad \boldsymbol{\Sigma}_\theta(t) = \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{I}$$

ELBO for Diffusion Model: Noise Reparametrization

- Recall the score matching term

$$D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \,||\, p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)\right) = D_{\mathrm{KL}}\left(\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) \,||\, \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)\right)$$
$$= \frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_\theta|}{|\boldsymbol{\Sigma}_q|} - d + \mathrm{tr}\left(\boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{\Sigma}_q\right) + \left(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q\right)^\top \boldsymbol{\Sigma}_\theta^{-1}\left(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q\right)\right]$$

- Plugging in the reparametrized mean and covariances

$$\sigma_q^2(t) = \begin{cases} \sigma_{\mathrm{rec}}^2, & t = 1 \\ \dfrac{(1-\alpha_t)(1-\overline{\alpha_{t-1}})}{1-\overline{\alpha_t}}, & t \geq 2 \end{cases}$$

$$= \frac{1}{2\sigma_q^2(t)}\left\|\boldsymbol{\mu}_q(\boldsymbol{x}_t, \boldsymbol{x}_0) - \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)\right\|_2^2$$
$$= \frac{1}{2\sigma_q^2(t)}\left\|\frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t, t) - \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t + \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\bar{\boldsymbol{\epsilon}}_t\right\|_2^2$$
$$= \frac{1}{2\sigma_q^2(t)}\frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_t)\alpha_t}\left\|\bar{\boldsymbol{\epsilon}}_t - \hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t, t)\right\|_2^2 = \frac{1}{2}\frac{(1-\alpha_t)}{(1-\overline{\alpha_{t-1}})\alpha_t}\left\|\bar{\boldsymbol{\epsilon}}_t - \hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t, t)\right\|_2^2$$

- Plugging into the ELBO we arrive to

$$\underset{\theta}{\mathrm{argmax}} - \sum_{t=1}^T \mathbb{E}_{q_\phi(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \,||\, p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)\right)\right] = \underset{\theta}{\mathrm{argmax}} - \sum_{t=1}^T \mathbb{E}_{q_\phi(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\frac{1}{2}\frac{(1-\alpha_t)}{(1-\overline{\alpha_{t-1}})\alpha_t}\left\|\bar{\boldsymbol{\epsilon}}_t - \hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t, t)\right\|_2^2\right]$$

DDPM Noise Reconstruction Loss

- The ELBO in terms of the noise is

$$\log p(\boldsymbol{x}) \geq \max_{\theta} \mathcal{L}_{\text{ELBO}}(\theta) = \max_{\theta} -\sum_{t=1}^{T} \mathbb{E}_{q_\phi(\boldsymbol{x}_t|\boldsymbol{x}_0)} \left[\frac{1}{2} \frac{(1-\alpha_t)}{(1-\overline{\alpha_{t-1}})\alpha_t} \|\bar{\epsilon}_t - \hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t, t)\|_2^2 \right]$$

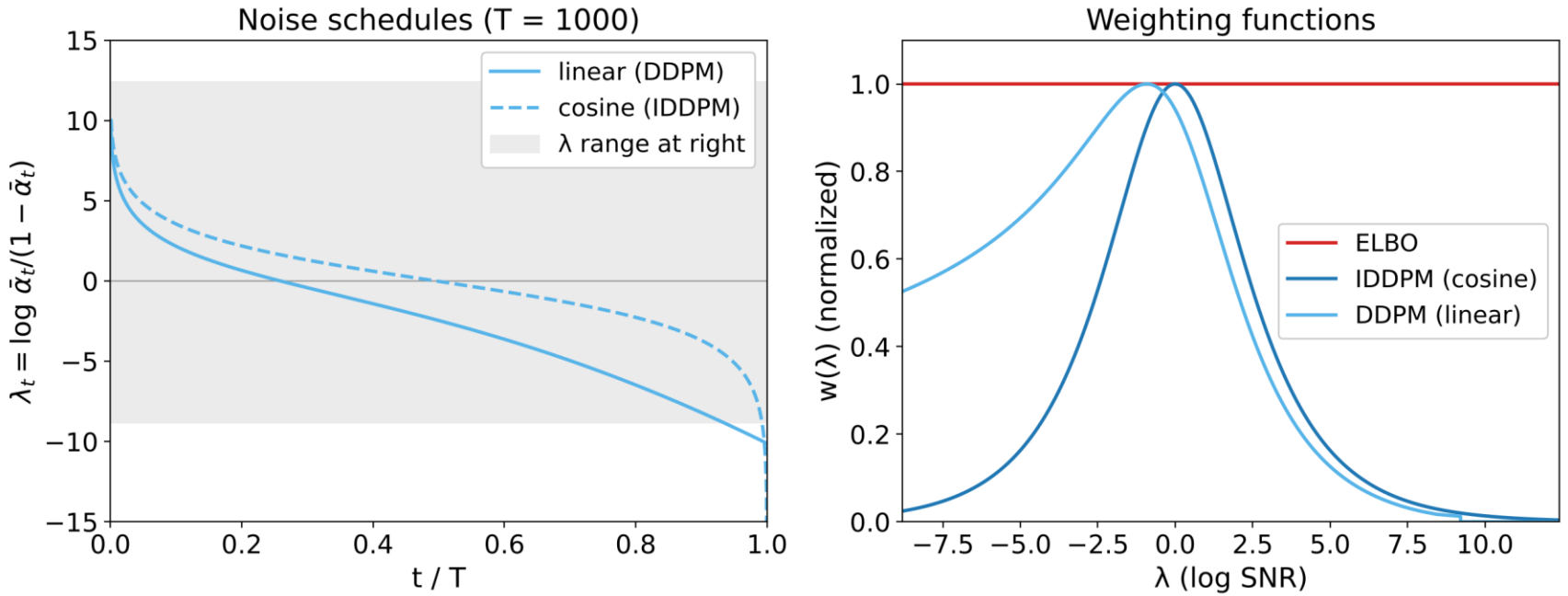
- However, the loss used in DDPM is

$$\max_{\theta} \mathcal{L}_{\text{DDPM}}(\theta) = \max_{\theta} -\sum_{t=1}^{T} \mathbb{E}_{q_\phi(\boldsymbol{x}_t|\boldsymbol{x}_0)} \left[\|\bar{\epsilon}_t - \hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t, t)\|_2^2 \right]$$

- What difference does it make to drop the weighting $w(t) = \frac{1}{2} \frac{(1-\alpha_t)}{(1-\overline{\alpha_{t-1}})\alpha_t}$?
  - The optimum is the same, $\underset{\theta}{\text{argmax}}\, \mathcal{L}_{\text{ELBO}}(\theta) = \underset{\theta}{\text{argmax}}\, \mathcal{L}_{\text{DDMP}}(\theta)$.
  - However, they differ in how much they penalize the error at the different levels of noise λ_t.

DDPM Noise Reconstruction Loss

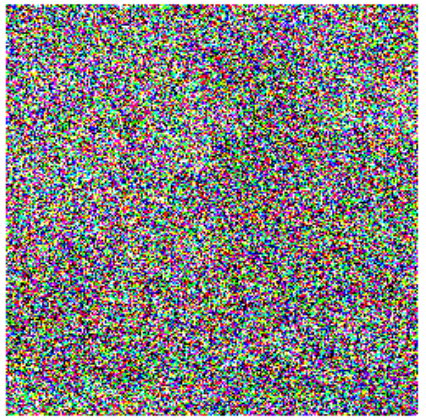
- Recall $\boldsymbol{x}_t = \sqrt{\overline{\alpha_t}}\boldsymbol{x}_0 + \sqrt{1-\overline{\alpha_t}}\,\overline{\boldsymbol{\epsilon}_t}$; define the Signal-to-Noise $\lambda_t = \log \overline{\alpha_t}/1-\overline{\alpha_t}$.



- ELBO assigns weight to each t according to how much Signal-to-Noise ratio it covers.

- DDPM assigns equal weight to each t.

The Reverse Diffusion Process for DDPM



- Putting it all together, the reverse diffusion process is given by:

$$p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) \sim \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(t))$$

- In DDPM, $\boldsymbol{\Sigma}_\theta(t)$ is not learned and is fixed to $\boldsymbol{\Sigma}_\theta(t) = (1 - \alpha_t)\,\boldsymbol{I}$, slightly larger than the matched variance. Both work well in practice.

- Therefore, we generate an image via the reverse diffusion process

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

where $\boldsymbol{x}_T \sim \mathcal{N}(\boldsymbol{x}_T; \boldsymbol{0}, \boldsymbol{I})$, $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{\epsilon}_t; \boldsymbol{0}, \boldsymbol{I})$ are i.i.d. and independent of $\boldsymbol{x}_t$.