

# Deep Generative Models: Diffusion Models

Fall Semester 2025

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

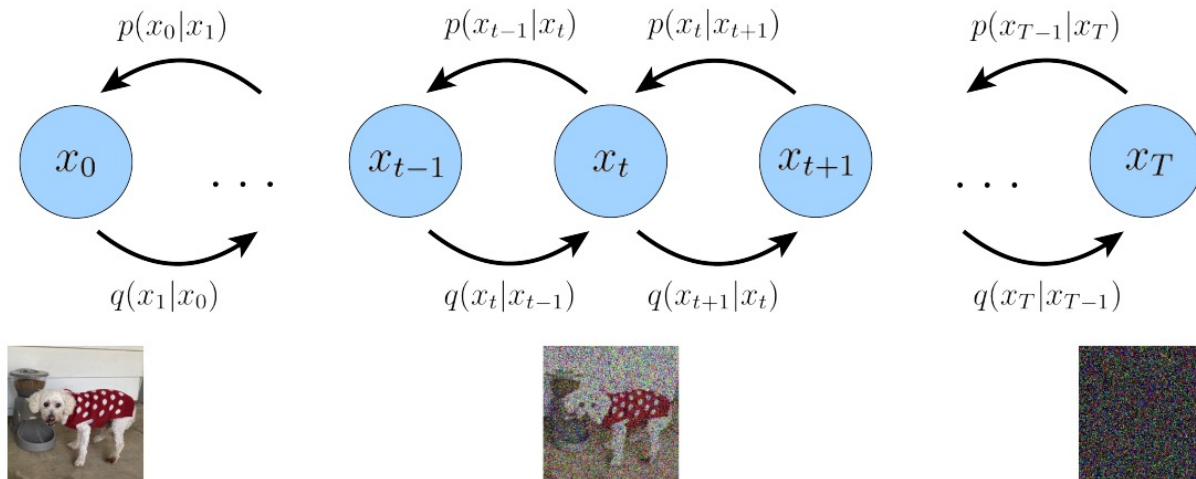
- Markov Hierarchical Variational Auto Encoders (MHVAEs)
  - Autoregressive Encoder and Autoregressive Decoder of an MHVAE
  - Derivation of the ELBO of an MHVAE
- **Diffusion Models as MHVAEs with a Linear Gaussian Autoregressive Latent Space**
  - Forward Diffusion Process
  - Reverse Diffusion Process
  - 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
  - Stable Diffusion: Text-Conditioned Diffusion Model
  - ControlNet: Multimodal Control for Consistent Synthesis

# Diffusion Model as MHVAEs with Gaussian Latents

- **A Diffusion Model is an MHVAE** where the latent variables  $x_{1:T}$  have the same dimension as the data  $x_0$ , and the encoder  $q_\phi(x_{1:T} | x_0) = \prod_{t=1}^T q_\phi(x_t | x_{t-1})$  is not learned, but it is pre-specified as a linear Gaussian model

$$q_\phi(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t} x_{t-1}, (1 - \alpha_t)I)$$
$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1 - \alpha_t} \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(\epsilon_t; 0, I)$$

- The parameter  $\alpha_t$  is chosen such that  $x_T \sim \mathcal{N}(x_T; 0, I)$  is a standard Gaussian



# The Forward Process of Diffusion Model

- Consider the formulation of a single noising step:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1 - \alpha_t} \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(\epsilon_t; 0, I).$$

- Note  $x_1 \mid x_0$  and  $x_t \mid x_{t-1}$  are Gaussian, hence  $x_t \mid x_0 \sim \mathcal{N}(x_t; \mu_q(x_0), \Sigma_q(x_0))$ .
- We can compute  $\mu_q(x_0) = \mathbb{E}[x_t \mid x_0]$ , recursively as follows:

$$\begin{aligned} \mathbb{E}[x_t \mid x_0] &= \mathbb{E}[\sqrt{\alpha_t} x_{t-1} + \sqrt{1 - \alpha_t} \epsilon_t \mid x_0] \\ &= \sqrt{\alpha_t} \mathbb{E}[x_{t-1} \mid x_0] + \sqrt{1 - \alpha_t} \mathbb{E}[\epsilon_t] \\ &= \sqrt{\alpha_t} \mathbb{E}[x_{t-1} \mid x_0] \\ &= \sqrt{\alpha_t} \sqrt{\alpha_{t-1}} \mathbb{E}[x_{t-2} \mid x_0] \\ &= \sqrt{\alpha_t} \sqrt{\alpha_{t-1}} \cdots \sqrt{\alpha_1} x_0 \\ &= \sqrt{\bar{\alpha}_t} x_0 \end{aligned}$$

$$\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$$

# The Forward Process of Diffusion Model

- We can compute  $\Sigma_q(x_0) = \text{Var}[x_t | x_0]$ , recursively as follows:

$$\begin{aligned}\text{Var}(x_t | x_0) &= \text{Var}(\sqrt{\alpha_t}x_{t-1} + \sqrt{1 - \alpha_t}\epsilon_t | x_0) \\ &= \alpha_t \text{Var}(x_{t-1} | x_0) + (1 - \alpha_t) \text{Var}(\epsilon_t) \\ &= \alpha_t \text{Var}(x_{t-1} | x_0) + (1 - \alpha_t) I\end{aligned}$$

- That is:

$$\begin{aligned}\text{Var}(x_t | x_0) &= \alpha_t [\alpha_{t-1} \text{Var}(x_{t-2} | x_0) + (1 - \alpha_{t-1}) I] + (1 - \alpha_t) I \\ &= \alpha_t \alpha_{t-1} \text{Var}(x_{t-2} | x_0) + (1 - \alpha_t \alpha_{t-1}) I \\ &= \dots \\ &= \alpha_t \alpha_{t-1} \dots \alpha_1 \text{Var}(x_0 | x_0) + \left(1 - \prod_{i=1}^t \alpha_i\right) I \\ &= \left(1 - \prod_{i=1}^t \alpha_i\right) I = (1 - \bar{\alpha}_t) I\end{aligned}$$

$$\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$$

# The Forward Process of Diffusion Model

- We have shown that  $x_t$  is a linear Gaussian transformation of  $x_0$  with scheduled randomness (controlled by  $\overline{\alpha}_t$ ) drawn from a standard normal distribution, i.e.,

$$x_t \mid x_0 \sim \mathcal{N}(x_t; \sqrt{\overline{\alpha}_t} x_0, (1 - \overline{\alpha}_t) I)$$

- Therefore, given  $x_0$ , we can sample  $x_t$  directly without having to generate all  $x_t$ 's:

$$x_t = \sqrt{\overline{\alpha}_t} x_0 + \sqrt{1 - \overline{\alpha}_t} \bar{\epsilon}_t, \quad \bar{\epsilon}_t \sim \mathcal{N}(\bar{\epsilon}_t; 0, I)$$

- Moreover, we can also generate  $x_0$  from  $x_t$  as

$$x_0 = (x_t - \sqrt{1 - \overline{\alpha}_t} \bar{\epsilon}_t) / \sqrt{\overline{\alpha}_t}, \quad \bar{\epsilon}_t \sim \mathcal{N}(\bar{\epsilon}_t; 0, I)$$

- This suggests we can reverse the noising process. However, exact reversal requires knowing the exact  $\bar{\epsilon}_t$ . The reverse diffusion process is designed to predict the noise  $\bar{\epsilon}_t$  that needs to be added to  $x_t$  to generate  $x_0$ .

# The Reverse Diffusion Process

- We have designed a forward diffusion process  $q_\phi(x_t | x_{t-1})$  that
  - At each step adds Gaussian noise to the input until it becomes pure noise
  - Allows us to sample  $x_t | x_0$  without having to compute  $x_t$  recursively
  - Allows us to sample  $x_0 | x_t$  without having to compute  $x_t$  recursively
- We now need to design a reverse diffusion process  $p_\theta(x_{t-1} | x_t)$  that makes the calculation of the ELBO easy. We do this by
  - Understanding the structure of  $q_\phi(x_t | x_{t-1})$
  - Making  $p_\theta(x_{t-1} | x_t)$  match that structure
- Recall the ELBO is given by:

$$= \underbrace{\mathbb{E}_{q_\phi(x_1|x_0)}[\log p_\theta(x_0 | x_1)]}_{\text{reconstruction term}} - \underbrace{D_{\text{KL}}(q_\phi(x_T | x_0) || p_\theta(x_T))}_{\text{prior matching term}} - \sum_{t=2}^T \underbrace{\mathbb{E}_{q_\phi(x_t|x_0)} [D_{\text{KL}}(q_\phi(x_{t-1} | x_t, x_0) || p_\theta(x_{t-1} | x_t))]}_{\text{score matching term}}$$

# ELBO for Diffusion Model: Score Matching Term

- To compute the third term, we need

$$q(x_{t-1} | x_t, x_0) = \frac{q(x_t | x_{t-1}, x_0) q(x_{t-1} | x_0)}{q(x_t | x_0)} = \frac{\mathcal{N}(x_t; \sqrt{\alpha_t} x_{t-1}, (1 - \alpha_t)I) \mathcal{N}(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}} x_0, (1 - \bar{\alpha}_{t-1})I)}{\mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t)I)}$$

$$\propto \exp\left(-\frac{\|x_t - \sqrt{\alpha_t} x_{t-1}\|^2}{2(1 - \alpha_t)}\right) \cdot \exp\left(-\frac{\|x_{t-1} - \sqrt{\bar{\alpha}_{t-1}} x_0\|^2}{2(1 - \bar{\alpha}_{t-1})}\right) \propto \exp\left(-\frac{\alpha_t \|x_{t-1} - \frac{x_t}{\sqrt{\alpha_t}}\|^2}{2(1 - \alpha_t)}\right) \cdot \exp\left(-\frac{\|x_{t-1} - \sqrt{\bar{\alpha}_{t-1}} x_0\|^2}{2(1 - \bar{\alpha}_{t-1})}\right)$$

- Applying the product rule, we get  $q(x_{t-1} | x_t, x_0) = \mathcal{N}(x_{t-1}; \mu_q(x_0), \Sigma_q)$ , where

$$\Sigma_q := \text{Cov}(x_{t-1} | x_t, x_0) = \left( \frac{\alpha_t}{1 - \alpha_t} I + \frac{1}{1 - \bar{\alpha}_{t-1}} I \right)^{-1} = \frac{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} I$$

$$\mu_q(x_t, x_0) := \mathbb{E}(x_{t-1} | x_t, x_0) = \Sigma_q \left( \frac{\alpha_t}{1 - \alpha_t} I \frac{x_t}{\sqrt{\alpha_t}} + \frac{1}{1 - \bar{\alpha}_{t-1}} I \sqrt{\bar{\alpha}_{t-1}} x_0 \right)$$

$$= \frac{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} I \left( \frac{\sqrt{\alpha_t}}{1 - \alpha_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_{t-1}} x_0 \right) = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})x_t + \bar{\alpha}_{t-1}(1 - \alpha_t)x_0}{1 - \bar{\alpha}_t}$$

$$\begin{aligned} q(x_t | x_{t-1}) &= \mathcal{N}(x_t; \sqrt{\alpha_t} x_{t-1}, (1 - \alpha_t)I) \\ q(x_t | x_0) &= \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t)I) \end{aligned}$$

$$\mathcal{N}(x; \mu_1, \Sigma_1) \mathcal{N}(x; \mu_2, \Sigma_2) \propto \mathcal{N}(x; \bar{\mu}, \bar{\Sigma})$$

$$\bar{\mu} = \bar{\Sigma} (\Sigma_1^{-1} \mu_1 + \Sigma_2^{-1} \mu_2), \bar{\Sigma} = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}$$

$$\begin{aligned} \mu_1 &= \frac{x_t}{\sqrt{\alpha_t}}, \Sigma_1 = \frac{1 - \alpha_t}{\alpha_t} I, \\ \mu_2 &= \sqrt{\bar{\alpha}_{t-1}} x_0, \Sigma_2 = (1 - \bar{\alpha}_{t-1}) I \end{aligned}$$



# ELBO for Diffusion Model: Decoder Matches Encoder

- Recall KL divergence for Gaussians

$$D_{\text{KL}} \left( \mathcal{N}(x; \mu_x, \Sigma_x) \parallel \mathcal{N}(y; \mu_y, \Sigma_y) \right) = \frac{1}{2} \left[ \log \frac{|\Sigma_y|}{|\Sigma_x|} - d + \text{tr}(\Sigma_y^{-1} \Sigma_x) + (\mu_y - \mu_x)^T \Sigma_y^{-1} (\mu_y - \mu_x) \right]$$

- Choosing mean of  $p_\theta(x_{t-1} \mid x_t)$  to match form of mean of  $q(x_{t-1} \mid x_t, x_0)$

$$\mu_q(x_t, x_0) = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)x_0}{1 - \bar{\alpha}_t} \Rightarrow \mu_\theta(x_t, t) = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)\widehat{x}_\theta(x_t, t)}{1 - \bar{\alpha}_t}$$

- Choosing variance of  $p_\theta(x_{t-1} \mid x_t)$  to match exactly variance of  $q(x_{t-1} \mid x_t, x_0)$

$$\Sigma_q = \frac{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} I \Rightarrow \Sigma_\theta = \frac{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} I$$

- The ELBO reduces to:

$$\sigma_q^2(t) = \frac{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}$$

$$\begin{aligned} D_{\text{KL}}(q(x_{t-1} \mid x_t, x_0) \parallel p_\theta(x_{t-1} \mid x_t)) &= D_{\text{KL}} \left( \mathcal{N}(x_{t-1}; \mu_q(x_t, x_0), \Sigma_q) \parallel \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_q) \right) \\ &= \frac{1}{2\sigma_q^2(t)} [\|\mu_\theta - \mu_q\|_2^2] = \frac{1}{2\sigma_q^2(t)} \frac{\bar{\alpha}_{t-1}(1 - \alpha_t)^2}{(1 - \bar{\alpha}_t)^2} [\|\widehat{x}_\theta(x_t, t) - x_0\|_2^2] \end{aligned}$$

# Reparameterization as an Alternative Form for ELBO

- Plugging  $x_0 = \frac{x_t - \sqrt{1 - \bar{\alpha}_t} \bar{\epsilon}_t}{\sqrt{\bar{\alpha}_t}}$  into the denoising transition mean  $\mu_q(x_t, x_0)$ , we obtain:

$$\begin{aligned}
 \mu_q(x_t, x_0) &= \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)x_0}{1 - \bar{\alpha}_t} \\
 &= \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t) \frac{x_t - \sqrt{1 - \bar{\alpha}_t} \bar{\epsilon}_t}{\sqrt{\bar{\alpha}_t}}}{1 - \bar{\alpha}_t} \\
 &= \frac{1 - \bar{\alpha}_t}{(1 - \bar{\alpha}_t)\sqrt{\alpha_t}} x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t} \sqrt{\alpha_t}} \bar{\epsilon}_t \\
 &= \frac{1}{\sqrt{\alpha_t}} x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t} \sqrt{\alpha_t}} \bar{\epsilon}_t
 \end{aligned}$$

- Choosing the mean and variance of  $p$  to match the mean and variance of  $q$ :

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t} \sqrt{\alpha_t}} \bar{\epsilon}_t, \quad \Sigma_\theta(t) = \frac{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} I$$

# The Reverse Diffusion Process for DDPM

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

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

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

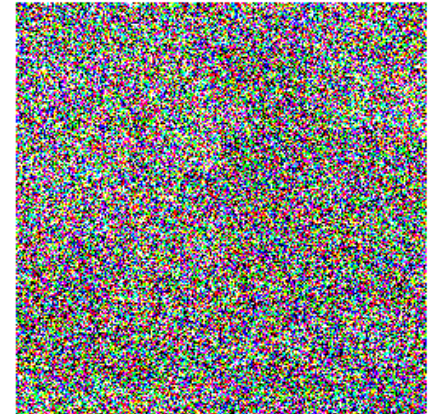
$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \hat{\epsilon}_{\theta}(x_t, t) \right) + \sqrt{\beta_t} \epsilon_t$$

- Where  $x_T \sim \mathcal{N}(x_T; 0, I)$ ,  $\epsilon_t \sim \mathcal{N}(\epsilon_t; 0, I)$ .

# Progressive Denoising or Direct Reconstruction?

- The model predicts the **noise** to be removed in each step by optimizing the **score matching term**. This reduces to minimizing the difference between the predicted noise and the ground-truth schedule noise:

$$\begin{aligned} & \underset{\theta}{\operatorname{argmin}} D_{\text{KL}}(q(x_{t-1} | x_t, x_0) \parallel p_{\theta}(x_{t-1} | x_t)) \\ &= \underset{\theta}{\operatorname{argmin}} D_{\text{KL}}(\mathcal{N}(x_{t-1}; \mu_q, \Sigma_q(t)) \parallel \mathcal{N}(x_{t-1}; \mu_{\theta}, \Sigma_q(t))) \\ &= \underset{\theta}{\operatorname{argmin}} \frac{1}{2\sigma_q^2(t)} \left[ \left\| \cancel{\frac{1}{\sqrt{\alpha_t}} x_t} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}} \hat{\epsilon}_{\theta}(x_t, t) - \cancel{\frac{1}{\sqrt{\alpha_t}} x_t} + \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}} \epsilon_0 \right\|_2^2 \right] \\ &= \underset{\theta}{\operatorname{argmin}} \frac{1}{2\sigma_q^2(t)} \frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_t)\alpha_t} [\|\epsilon_0 - \hat{\epsilon}_{\theta}(x_t, t)\|_2^2] \end{aligned}$$



- Predicting  $x_0$  from a highly noisy  $x_t$  in one step is complex, as the signal is dominated by significant noise for large  $t$ .
- Predicting the noise at each step and refining  $x_t$  towards  $x_0$  makes learning more manageable (e.g., it converges faster or it requires a smaller network capacity).