# Lecture-02 VAE, DDPM, Discrete-VAE, D3PMs

① AE – Autoencoder

② VAE – Variational Autoencoder

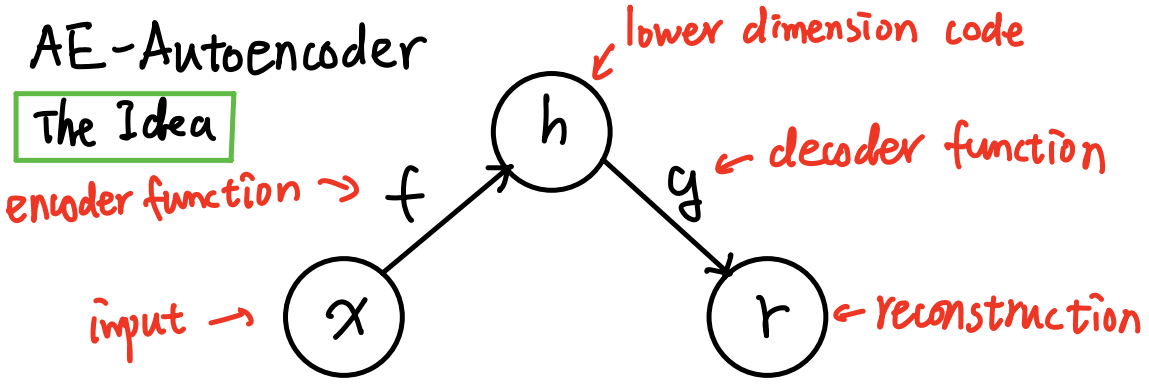
③ DDPM – Denoising Diffusion Probabilistic Model

④ Discrete – VAE

⑤ D3PMs

## ①. AE-Autoencoder

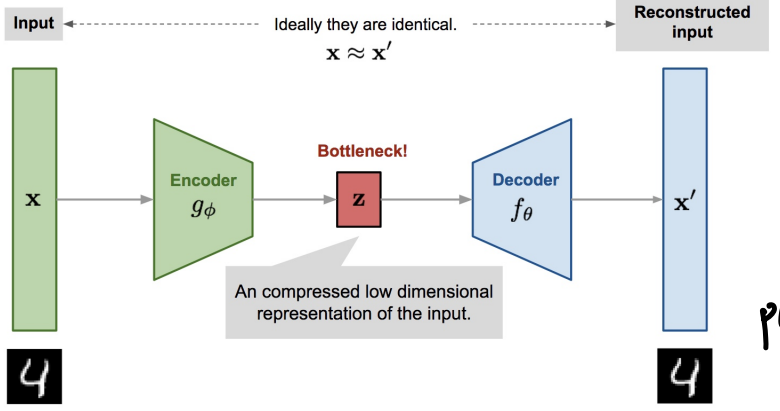
- The Idea




training : attempt to copy its input to its output

$r = g(f(x)) \Rightarrow \min_{f,g} L(x, g(f(x)))$

application : dimension reduction, feature learning




- Encoder: $x \rightarrow z$
- Decoder: $z \rightarrow x'$

parameters = $[\phi, \theta]$

- Ideas of autoencoder originated in 1980s, and later promoted by [Hinton and Salakhutdinov, 2006]
- Application to dimensionality reduction, like PCA
- If $x$ are images: $(x \approx f_\theta(g_\phi(x)))$

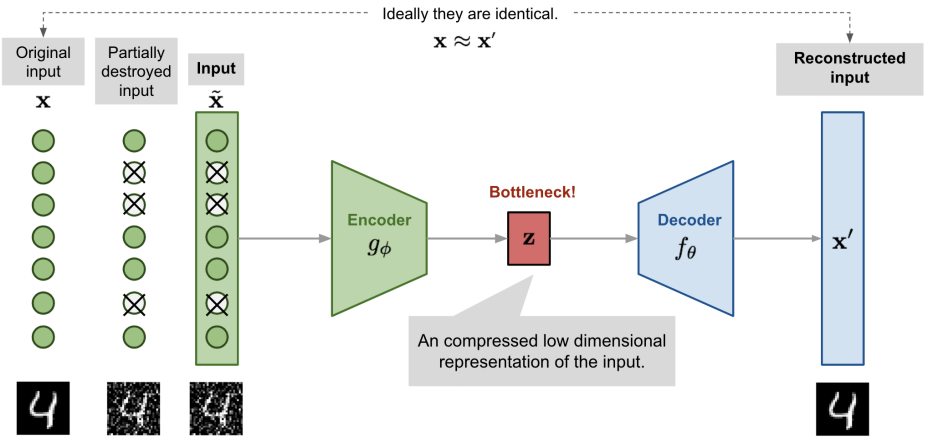
$$L_{AE}(\theta,\phi) = \frac{1}{N}\sum_{i=1}^{N}\left(x^{(i)} - f_\theta(g_\phi(x^{(i)}))\right)^2$$

If $x$ are categorical:

$$L_{AE}(\theta,\phi) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k}\left(-x_k^{(i)} \cdot \log \hat{x}_k^{(i)}\right),$$

$\hat{x}_k^{(i)}$ is the output of AE. } AE: Easy overfitting

- Many variants:
  - Denoising autoencoder: Add noise to $x^{(i)}$




$$L_{DAE}(\theta,\phi) = \frac{1}{N}\sum_{i=1}^{N}\left(x^{(i)} - f_\theta(g_\theta(\tilde{x}^{(i)}))\right)^2,$$

$$\tilde{x}^{(i)} \sim M_D(\tilde{x}^{(i)} \mid x^{(i)})$$

- Sparse AE
- k-Sparse AE
- Contractive AE

Key limitations:
The latent space of AEs is unstructured: randomly sampling $z$ produces meaningless outputs.

① AE – Autoencoder

② VAE – Variational Autoencoder

③ DDPM – Denoising Diffusion Probabilistic Model

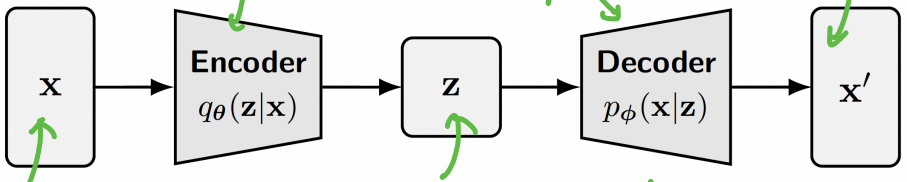
④ Discrete – VAE

⑤ D3PMs

- New Idea : [Kingma and Welling, 2013]

Auto-Encoding Variational Bayes

From $z \rightarrow p(z)$ : imposing a probabilistic structure on the latent space. → a true generative model.




observed variables
data point (image)

latent variables
capture object shape, color, or style.

- generation of $x'$ :

Step 1: $z \sim p_{prior} := N(0, I)$

Step 2: decoding $z$ via $p_\phi$ : $x' \sim p_\phi(x|z)$

latent-variable generative model :

$$p_\phi(x) = \int p_\phi(x|z) \cdot p(z)\, dz \Rightarrow \text{goal: } \max_\phi p_\phi(x)$$

intractable, but can use variational method.

- construction of encoder (inference network)

  given $x$, what latent code $z$ could have produced it?

$$p_\phi(z|x) = \frac{p_\phi(x|z)\cdot p(z)}{p_\phi(x)} \leftarrow \text{intractable}$$

posterior

Key: "variational" step in VAEs, replacing the intractable posterior with a tractable approximation!

Encoder $q_\theta(z|x)$: learnable proxy.

$$q_\theta(z|x) \simeq p_\phi(z|x).$$

- Training via ELBO

**Theorem.1 Evidence Lower Bound (ELBO)**

For any data point $x$, the log-likelihood satisfies:

$$\log p_\phi(x) \geq \mathcal{L}_{ELBO}(\theta, \phi; x),$$

$$\mathcal{L}_{ELBO} = \underbrace{\mathbb{E}_{z\sim q_\theta(z|x)}[\log p_\phi(x|z)]}_{\text{Reconstruction term}} - \underbrace{D_{KL}(q_\theta(z|x)\,\|\,p(z))}_{\text{Latent Regularization}}$$

Proof:

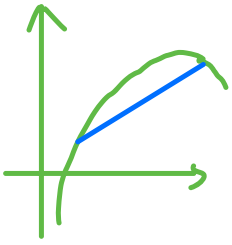
$$\log p_\phi(x) = \log \int p_\phi(x,z)\,dz$$

$$= \log \int q_\theta(z|x)\cdot \frac{p_\phi(x,z)}{q_\theta(z|x)}\,dz$$

$$= \log \mathbb{E}_{z\sim q_\theta(z|x)}\left[\frac{p_\phi(x,z)}{q_\theta(z|x)}\right]$$

Jensen's Inequality →

$$\geq \mathbb{E}_{z\sim q_\theta(z|x)}\left[\log \frac{p_\phi(x,z)}{q_\theta(z|x)}\right]$$



Note $P_\phi(x,z) = P_\phi(x|z) \cdot P(z)$

$$\log P_\phi(x) \geq \mathbb{E}_{z\sim q_\theta(z|x)}\left[\log \frac{P_\phi(x|z)\cdot P(z)}{q_\theta(z|x)}\right]$$

$$= \mathbb{E}_{z\sim q_\theta}[\log P_\phi(x|z)] - D_{KL}(q_\theta(z|x) \,||\, P(z))$$

Reconstruction: encourages accurate recovery of $x$ from $z$

Latent KL: encourages the encoder $q_\theta(z|x)$ to stay close to $P_{prior}(z)$ → enable meaningful generation.

- Information-theoretic Interpretation

Denote
- generative joint: $P_\phi(x,z) = P(z)\cdot P_\phi(x|z)$
- inference joint: $q_\theta(x,z) = P_{data}(x)\cdot q_\theta(z|x)$

Recall our goal: $D_{KL}(P_{data}(x) \,||\, P_\phi(x))$

We have:

$D_{KL}(P_{data}(x) \,||\, P_\phi(x)) \leq D_{KL}(q_\theta(x,z) \,||\, P_\phi(x,z))$

↑ chain Rule for KL Divergence

Proof:

$D_{KL}(q_\theta(x,z) \,||\, P_\phi(x,z))$ : Total error bound

$$= \mathbb{E}_{q_\theta(x,z)}\left[\log \frac{P_{data}(x)\cdot q_\theta(z|x)}{P_\phi(x)\cdot P_\phi(z|x)}\right]$$

$$= \mathbb{E}_{P_{data}(x)}\left[\log \frac{P_{data}(x)}{P_\phi(x)} + D_{KL}(q_\theta(z|x) \,||\, P_\phi(z|x))\right]$$

// Note : $\mathbb{E}_{q_\theta(x,z)} = \mathbb{E}_{q_\theta(z|x)\cdot p_{data}(x)}$

$= \underbrace{D_{KL}(p_{data} \| p_\phi)}_{\text{True modeling error} \geq 0} + \underbrace{\mathbb{E}_{p_{data}(x)}[D_{KL}(q_\theta(z|x) \| p_\phi(z|x))]}_{\text{Inference error} \geq 0}$

---

$$\log p_\phi(x) = \log \mathbb{E}_{z\sim q_\theta(z|x)}\left[\frac{p_\phi(x,z)}{q_\theta(z|x)}\right]$$

$$L_{ELBO}(\phi,\theta;x) = \mathbb{E}_{z\sim q_\theta}[\log p_\phi(x|z)] - D_{KL}(q_\theta(z|x) \| p(z))$$

$$\log p_\phi(x) - L_{ELBO}(\phi,\theta;x) = D_{KL}(q_\theta(z|x) \| p_\phi(z|x))$$

↑ gap between log-likelihood and the ELBO proxy.

↑ inference error.

maximizing $L_{ELBO}$ => reducing inference error.

Proof:

$$\begin{aligned}
\mathcal{D}_{\mathrm{KL}}(q_\theta(z \mid x) \| p_\phi(z \mid x)) &= \int q_\theta(z \mid x) \log \frac{q_\theta(z \mid x)}{p_\phi(z \mid x)} \mathrm{d}z \\
&= \int q_\theta(z \mid x) \log \frac{q_\theta(z \mid x) p_\phi(x)}{p_\phi(z, x)} \mathrm{d}z \\
&= \int q_\theta(z \mid x) \left(\log p_\phi(x) + \log \frac{q_\theta(z \mid x)}{p_\phi(z, x)}\right) \mathrm{d}z \\
&\stackrel{*}{=} \log p_\phi(x) + \int q_\theta(z \mid x) \log \frac{q_\theta(z \mid x)}{p_\phi(z, x)} \mathrm{d}z \\
&= \log p_\phi(x) + \int q_\theta(z \mid x) \log \frac{q_\theta(z \mid x)}{p_\phi(x \mid z) p_\phi(z)} \mathrm{d}z \\
&= \log p_\phi(x) + \mathbb{E}_{z \sim q_\theta(z|x)}\left[\log \frac{q_\theta(z \mid x)}{p_\phi(z)} - \log p_\phi(x \mid z)\right] \\
&= \log p_\phi(x) + \mathcal{D}_{\mathrm{KL}}(q_\theta(z \mid x) \| p_\phi(z)) - \mathbb{E}_{z \sim q_\theta(z|x)}[\log p_\phi(x \mid z)],
\end{aligned}$$

where $\stackrel{*}{=}$ is due to $\int q_\theta(z \mid x) \mathrm{d}z = 1$.

$= \log p_\phi(x) - L_{ELBO}(\phi,\theta;x)$

## Gaussian VAE

- **Encoder**: $q_\theta(z|x)$ is modeled as a Gaussian

$$q_\theta(z|x) := \mathcal{N}(z; \mu_\theta(x), \mathrm{diag}(\sigma_\theta^2(x))),$$

with $\mu_\theta : \mathbb{R}^D \to \mathbb{R}^d$, $\sigma_\theta : \mathbb{R}^D \to \mathbb{R}_+^d$

- **Decoder**: modeled as Gaussian

$$p_\phi(x|z) := \mathcal{N}(x; \mu_\phi(z), \sigma^2 I),$$

with $\mu_\phi(z) : \mathbb{R}^d \to \mathbb{R}^D$ : NN

Under this model assumption

$$L_{ELBO} \begin{cases} \text{Reconstruction term: } \mathbb{E}_{q_\theta(z|x)}[\log p_\phi(x|z)] \\ \text{Latent Regularization: } -D_{KL}(q_\theta(z|x) \| p_{prior}(z)) \end{cases}$$

- **Reconstruction**:

$$\mathbb{E}_{q_\theta(z|x)}[\log p_\phi(x|z)] = -\frac{1}{2\sigma^2}\mathbb{E}_{q_\theta(z|x)}[\|x - \mu_\theta(z)\|^2] + C$$

⇓ ELBO objective

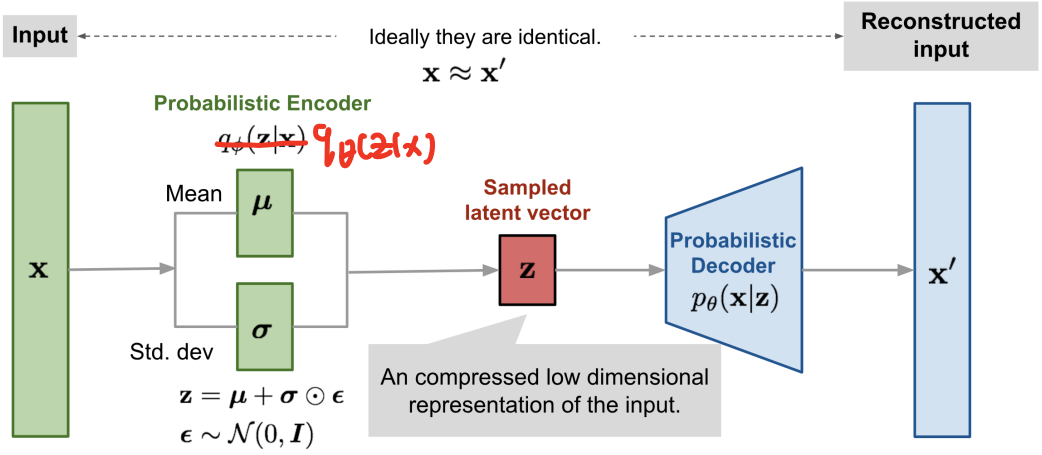
$$\min_{\theta,\phi} \mathbb{E}_{q_\theta(z|x)}\left[\frac{1}{2\sigma^2}\|x - \mu_\theta(z)\|^2\right] + D_{KL}(q_\theta(z|x) \| p_{prior}(z))$$

**Proof:** For D-dimensional Gaussian with mean $\mu$ and covariance $\sigma^2 I$:

$$p(x) = \frac{1}{(2\pi\sigma^2)^{D/2}} \cdot \exp\left(-\frac{1}{2\sigma^2}\|x - \mu\|^2\right)$$

$$\log p(x) = -\frac{1}{2\sigma^2}\|x - \mu\|^2 - \frac{D}{2}\log(2\pi\sigma^2).$$

Replacing $\mu$ with $\mu_\theta(z)$ and $C = -\frac{D}{2}\log(2\pi\sigma^2)$ ■




- $q_\theta(z|x)$ For sample $x^{(i)}$ :

$$z \sim q_\theta(z|x^{(i)}) = \mathcal{N}(z; \mu_i, \sigma_i^2 \cdot I)$$

$$z = \mu + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

- **Drawbacks of Standard VAE**

VAEs often produces blurry outputs

Consider: { fixed Gaussian encoder $q_{enc}(z|x)$
decoder $p_{dec}(x|z) = \mathcal{N}(x; \mu(z), \sigma^2 I)$

$$\Rightarrow \text{ELBO}: \arg\min_{\mu} \mathbb{E}_{p_{data}(x) \cdot q_{enc}(z|x)} [\|x - \mu(z)\|^2]$$

$$\Rightarrow \mu^*(z) = \mathbb{E}_{q_{enc}(x|z)}[x] \quad \text{with}$$

$$q_{enc}(x|z) = \frac{q_{enc}(z|x) \cdot p_{data}(x)}{p_{prior}(z)}$$

$$\Rightarrow \mu^*(z) = \frac{\mathbb{E}_{p_{data}(x)}[q_{enc}(z|x) \cdot x]}{\mathbb{E}_{p_{data}(x)}[q_{enc}(z|x)]}$$

↑ missing proof ?

If two inputs $x \neq x'$, such that $q_{enc}(\cdot|x)$ $q_{enc}(\cdot|x')$ overlap
⇒ share $z$ → blurry

Proof: $\mathbb{E}_{x,z}[f(x,z)] = \mathbb{E}_z[\mathbb{E}_{x|z}[f(x,z)]]$,

with $f(x,z) = \|x - \mu(z)\|^2$

$\mathbb{E}_{p(x)}\mathbb{E}_{q(z|x)}[\cdot] = \mathbb{E}_{q(z)}\mathbb{E}_{q(x|z)}[\cdot]$

$L(\mu) = \mathbb{E}_{q(z)}\left[\mathbb{E}_{q(x|z)}[\|x - \mu(z)\|^2]\right]$

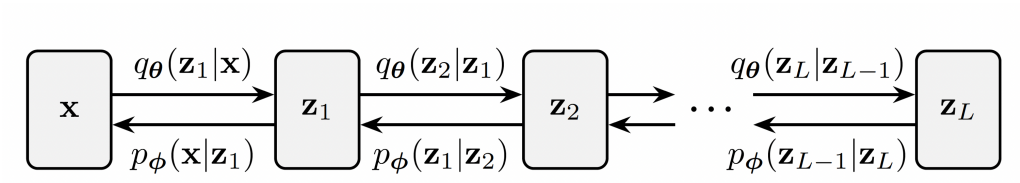
$\arg\min_{\mu} L(\mu) = \arg\min_{\mu} \mathbb{E}_{q(x|z)}[\|x - \mu(z)\|^2]$

Each subproblem:

$$\min_{\mu(z)} \mathbb{E}_{q(x|z)}[\|x - \mu(z)\|^2] \Leftrightarrow \mu^*(z) = \mathbb{E}_{q(x|z)}[x]$$

---

- From VAE to Hierarchical VAEs




HVAEs introduce multiple layers of latent code $z$

$$p_\phi(x, z_{1:L}) = p_\phi(x|z_1) \cdot \prod_{i=1}^{L} p_\phi(z_{i-1}|z_i) \cdot p(z_L)$$

$$p_{HVAE}(x) := \int p_\phi(x, z_{1:L})\, dz_{1:L}$$

$$q_\theta(z_{1:L}|x) = q_\theta(z_1|x) \cdot \prod_{i=2}^{L} q_\theta(z_i|z_{i-1})$$

HVAE's ELBO:

$$
\begin{aligned}
\log p_{\mathrm{HVAE}}(\mathbf{x}) &= \log \int p_{\phi}(\mathbf{x}, \mathbf{z}_{1:L}) \, \mathrm{d}\mathbf{z}_{1:L} \\
&= \log \int \frac{p_{\phi}(\mathbf{x}, \mathbf{z}_{1:L})}{q_{\theta}(\mathbf{z}_{1:L}|\mathbf{x})} q_{\theta}(\mathbf{z}_{1:L}|\mathbf{x}) \, \mathrm{d}\mathbf{z}_{1:L} \\
&= \log \mathbb{E}_{q_{\theta}(\mathbf{z}_{1:L}|\mathbf{x})} \left[ \frac{p_{\phi}(\mathbf{x}, \mathbf{z}_{1:L})}{q_{\theta}(\mathbf{z}_{1:L}|\mathbf{x})} \right] \\
&\geq \mathbb{E}_{q_{\theta}(\mathbf{z}_{1:L}|\mathbf{x})} \left[ \log \frac{p_{\phi}(\mathbf{x}, \mathbf{z}_{1:L})}{q_{\theta}(\mathbf{z}_{1:L}|\mathbf{x})} \right] \\
&=: \mathcal{L}_{\mathrm{ELBO}}(\phi).
\end{aligned}
$$

Substituting the factorized forms yields:

$$
\mathcal{L}_{\mathrm{ELBO}} = \mathbb{E}_{q_{\theta}(\mathbf{z}_{1:L}|\mathbf{x})} \left[ \log \frac{p(\mathbf{z}_L) \prod_{i=2}^{L} p_{\phi}(\mathbf{z}_{i-1}|\mathbf{z}_i) p_{\phi}(\mathbf{x}|\mathbf{z}_1)}{q_{\theta}(\mathbf{z}_1|\mathbf{x}) \prod_{i=2}^{L} q_{\theta}(\mathbf{z}_i|\mathbf{z}_{i-1})} \right].
$$

**Observation 2 (Key observations from HVAE)**

*Stacking layers allows the model to generate data progressively, starting with coarse details and adding finer ones at each step. This process makes it far easier to capture the complex structure of high-dimensional data.*

Figure 1: 256×256-pixel samples generated by NVAE, trained on CelebA HQ [28].

- Deeper Networks in a flat VAE are not enough.

① variational family is limited:

- Greater network depth improves $\mu_\theta$ and $\sigma_\theta$ but does not expand the family
- if the decoder is too expensive, the model may suffer from posterior collapse

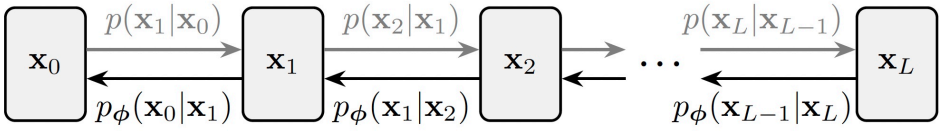
① AE – Autoencoder

② VAE – Variational Autoencoder

③ DDPM – Denoising Diffusion Probabilistic Model

④ Discrete – VAE

⑤ D3PMs

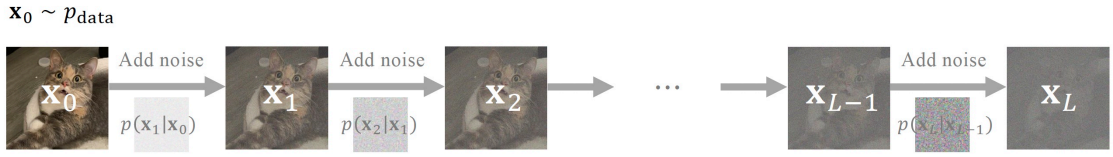



DDPMs { ① Forward Pass ( Fixed Encoder )
② Reverse Denoising Process ( Learnable Decoder )

- DDPMs achieve remarkable stability and expressive power.

○ Forward Process ( Fixed Encoder )

$x_0 \sim p_{data} \longrightarrow x_L \sim p_{prior} := N(0, I)$




Each step is governed by a fixed Gaussian transitional kernel

$$p(x_i|x_{i-1}) := N(x_i; \sqrt{1-\beta_i^2}\, x_{i-1}, \beta_i^2 I),$$

$i = 1, 2, \cdots, L$ with $x_0 \sim p_{data}$.

$\{\beta_i\}_{i=1}^{L}$ : pre-determined or learned. $\beta_i \uparrow$ and $\beta_i \in (0,1)$

Define $\alpha_i := \sqrt{1-\beta_i^2}$. Based on the definition of $p(x_i|x_{i-1})$,

$$x_i = \alpha_i \cdot x_{i-1} + \beta_i \cdot \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, I)$$

↖ controlled noise

[This is from the fact of sampling from a Gaussian
If $X \sim N(\mu, \sigma^2 I)$, the the sample can be generated as: $X = \mu + \sigma \cdot \varepsilon$, $\varepsilon \sim N(0, I)$. ]

Applying the transition kernels recursively:

$$x_1 = \alpha_1 x_0 + \beta_1 \varepsilon_1$$

$$x_2 = \alpha_2 x_1 + \beta_2 \varepsilon_2$$
$$= \alpha_2(\alpha_1 x_0 + \beta_1 \varepsilon_1) + \beta_2 \varepsilon_2$$
$$= \alpha_2 \alpha_1 x_0 + \alpha_2 \beta_1 \varepsilon_1 + \beta_2 \varepsilon_2$$

$$x_3 = \alpha_3 x_3 + \beta_3 \varepsilon_3$$
$$\vdots$$
$$= \alpha_3 \alpha_2 \alpha_1 x_0 + \alpha_3 \alpha_2 \beta_1 \varepsilon_1 + \alpha_3 \beta_2 \varepsilon_2 + \beta_3 \cdot \varepsilon_3$$

$$\Rightarrow x_i = \left(\prod_{j=1}^{i} \alpha_j\right) \cdot x_0 + \sum_{k=1}^{i} \left(\beta_k \cdot \prod_{j=k+1}^{i} \alpha_j\right) \cdot \varepsilon_k$$

Note, they are i.i.d.

Define $\bar{\alpha}_i = \prod_{j=1}^{i} \alpha_j$

$$= \sum_{k=1}^{i} \beta_k \cdot \frac{\bar{\alpha}_i}{\alpha_k}$$

To show $\mathrm{Var}\left[\sum_{k=1}^{i} \beta_k \cdot \frac{\bar{\alpha}_i}{\alpha_k} \varepsilon_k\right] = 1 - \bar{\alpha}_i^2$

Note $1 - \bar{\alpha}_i^2 = 1 - \prod_{j=1}^{i} \alpha_j^2$ and

telescoping ↙

$$\begin{cases} 1 - \alpha_i^2 = \beta_i^2 \\ 1 - \alpha_i^2 \alpha_{i-1}^2 = (1 - \alpha_i^2) + (1 - \alpha_{i-1}^2) \cdot \alpha_i^2 \\ 1 - \alpha_i^2 \alpha_{i-1}^2 \alpha_{i-2}^2 = (1 - \alpha_i^2) + (1 - \alpha_{i-1}^2) \cdot \alpha_i^2 \\ \qquad + (1 - \alpha_{i-2}^2) \cdot \alpha_{i-1}^2 \alpha_i^2 \end{cases}$$

variance ↙

$$\sum_{k=1}^{i} \left(\beta_k \cdot \frac{\bar{\alpha}_i}{\alpha_k}\right)^2 = \sum_{k=1}^{i} (1 - \alpha_k^2) \cdot \frac{\bar{\alpha}_i^2}{\alpha_k^2}$$

So, the mean is $\bar{\alpha}_i \cdot x_0$ and variance is $(1 - \bar{\alpha}_i^2) \cdot I$

Perturbation Kernel and Prior Distribution:

$$p_i(x_i|x_0) = \mathcal{N}(x_i; \bar{\alpha}_i \cdot x_0, (1-\bar{\alpha}_i^2)\cdot I),$$

$$\bar{\alpha}_i = \prod_{k=1}^{i} \alpha_k = \prod_{k=1}^{i} \sqrt{1-\beta_k^2} = \prod_{k=1}^{i} \alpha_k$$

Sampling: $x_i = \bar{\alpha}_i \cdot x_0 + \sqrt{1-\bar{\alpha}_i^2} \cdot \epsilon, \ \epsilon \sim \mathcal{N}(0, I)$

Since $\{\beta_i\}_{i=1}^{L}$ ↑ and $\beta_i \in (0,1)$, $\beta_L \to 1$.

$\Rightarrow \sqrt{1-\beta_L^2} \to 0$, $\bar{\alpha}_i \to 0 \Rightarrow \sqrt{1-\bar{\alpha}_i^2} \to 1$.

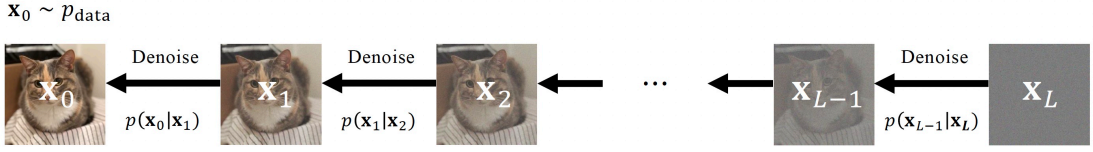
$p_L(x_L|x_0) \to \mathcal{N}(0, I)$ as $L \to \infty$,

which motivates the choice of the $p_{prior}$ as:

$p_{prior} := \mathcal{N}(0, I)$ ↔ No reliance on data $x_0$.

- **Reverse Denoising Process (Learnable Decoder)**

$x_L \sim p_{prior} \xrightarrow{\text{denoise}} x_0$ : meaningful data point




Key Q: Can we effectively approximate these reverse transition kernels $p(x_{i-1}|x_i)$ especially when $x_i \sim p_i(x_i)$ is complex?

To approximate $p(x_{i-1}|x_i)$, we use $p_\phi(x_{i-1}|x_i)$ and minimize

(*) $$\min_\phi \mathbb{E}_{p_i(x_i)}[D_{KL}(p(x_{i-1}|x_i) \,||\, p_\phi(x_{i-1}|x_i))]$$

Note $p(x_{i-1}|x_i) = p(x_i|x_{i-1}) \cdot \frac{p_{i-1}(x_{i-1})}{p_i(x_i)}$ ← intractable

as $p_t(x_t) = \int p_t(x_t|x_0) \cdot p_{data}(x_0)\, dx_0$ — no closed-form

Key Idea: we condition the reverse transition on a clean data sample $x$. → make the kernel tractable.

$$\Rightarrow \quad p(x_{t-1}|x_t, x) = \frac{p(x_t|x_{t-1}, x)}{p(x_t|x_{t-1})} \cdot \frac{p(x_{t-1}|x)}{p(x_t|x)}$$

(Markov property of $p(x_t|x_{t-1}, x) = p(x_t|x_{t-1})$ from the forward process assumption)

(*) => [ Marginal KL Minimization ⇔ Conditional KL Min.]

**Theorem 2.2.1: Equivalence Between Marginal and Conditional KL Minimization**

The following equality holds:

$$\begin{aligned} &\mathbb{E}_{p_i(\mathbf{x}_i)}\left[\mathcal{D}_{\mathrm{KL}}(p(\mathbf{x}_{i-1}|\mathbf{x}_i)\|p_\phi(\mathbf{x}_{i-1}|\mathbf{x}_i))\right] \\ &= \mathbb{E}_{p_{\mathrm{data}}(\mathbf{x})}\mathbb{E}_{p(\mathbf{x}_i|\mathbf{x})}\left[\mathcal{D}_{\mathrm{KL}}(p(\mathbf{x}_{i-1}|\mathbf{x}_i, \mathbf{x})\|p_\phi(\mathbf{x}_{i-1}|\mathbf{x}_i))\right] + C, \end{aligned} \tag{2.2.3}$$

where $C$ is a constant independent of $\phi$. Moreover, the minimizer of Equation (2.2.3) satisfies

$$p^*(\mathbf{x}_{i-1}|\mathbf{x}_i) = \mathbb{E}_{p(\mathbf{x}|\mathbf{x}_i)}\left[p(\mathbf{x}_{i-1}|\mathbf{x}_i, \mathbf{x})\right] = p(\mathbf{x}_{i-1}|\mathbf{x}_i), \quad \mathbf{x}_i \sim p_i.$$

***Proof.*** **Derivation of Equation (2.2.3).**

We start by expanding the right-hand side expectation:

$$\begin{aligned} &\mathbb{E}_{p(\mathbf{x}_0,\mathbf{x}_i)}\left[\mathcal{D}_{\mathrm{KL}}(p(\mathbf{x}_{i-1}|\mathbf{x}_i, \mathbf{x}_0)\|p_\phi(\mathbf{x}_{i-1}|\mathbf{x}_i))\right] \\ &= \int\int p(\mathbf{x}_0, \mathbf{x}_i)\mathcal{D}_{\mathrm{KL}}(p(\mathbf{x}_{i-1}|\mathbf{x}_i, \mathbf{x}_0)\|p_\phi(\mathbf{x}_{i-1}|\mathbf{x}_i))\, d\mathbf{x}_0\, d\mathbf{x}_i. \end{aligned}$$

By the definition of KL divergence,

$$\mathcal{D}_{\mathrm{KL}}(p(\mathbf{x}_{i-1}|\mathbf{x}_i, \mathbf{x}_0)\|p_\phi(\mathbf{x}_{i-1}|\mathbf{x}_i)) = \int p(\mathbf{x}_{i-1}|\mathbf{x}_i, \mathbf{x}_0) \log \frac{p(\mathbf{x}_{i-1}|\mathbf{x}_i, \mathbf{x}_0)}{p_\phi(\mathbf{x}_{i-1}|\mathbf{x}_i)}\, d\mathbf{x}_{i-1}.$$

Substituting this into the expectation, we have

$$\int\int\int p(\mathbf{x}_0, \mathbf{x}_i)p(\mathbf{x}_{i-1}|\mathbf{x}_i, \mathbf{x}_0) \log \frac{p(\mathbf{x}_{i-1}|\mathbf{x}_i, \mathbf{x}_0)}{p_\phi(\mathbf{x}_{i-1}|\mathbf{x}_i)}\, d\mathbf{x}_{i-1}\, d\mathbf{x}_0\, d\mathbf{x}_i.$$

Using the chain rule of probability,

$$p(\mathbf{x}_0, \mathbf{x}_i) = p(\mathbf{x}_i)p(\mathbf{x}_0|\mathbf{x}_i),$$

we rewrite the integral as

$$\int p(\mathbf{x}_i)\int p(\mathbf{x}_0|\mathbf{x}_i)\int p(\mathbf{x}_{i-1}|\mathbf{x}_i, \mathbf{x}_0) \log \frac{p(\mathbf{x}_{i-1}|\mathbf{x}_i, \mathbf{x}_0)}{p_\phi(\mathbf{x}_{i-1}|\mathbf{x}_i)}\, d\mathbf{x}_{i-1}\, d\mathbf{x}_0\, d\mathbf{x}_i.$$

This allows us to express the expectation in nested form:

$$\mathbb{E}_{p(\mathbf{x}_i)}\left[\mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_i)}\left[\mathbb{E}_{p(\mathbf{x}_{i-1}|\mathbf{x}_i,\mathbf{x}_0)}\left[\log\frac{p(\mathbf{x}_{i-1}|\mathbf{x}_i,\mathbf{x}_0)}{p_\phi(\mathbf{x}_{i-1}|\mathbf{x}_i)}\right]\right]\right].$$

Next, we apply the decomposition of the logarithm:

$$\log\frac{p(\mathbf{x}_{i-1}|\mathbf{x}_i,\mathbf{x}_0)}{p_\phi(\mathbf{x}_{i-1}|\mathbf{x}_i)} = \log\frac{p(\mathbf{x}_{i-1}|\mathbf{x}_i,\mathbf{x}_0)}{p(\mathbf{x}_{i-1}|\mathbf{x}_i)} + \log\frac{p(\mathbf{x}_{i-1}|\mathbf{x}_i)}{p_\phi(\mathbf{x}_{i-1}|\mathbf{x}_i)}.$$

Substituting this back into the expectation gives two terms:

$$\begin{aligned}&\mathbb{E}_{p(\mathbf{x}_i)}\left[\mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_i)}\left[\mathbb{E}_{p(\mathbf{x}_{i-1}|\mathbf{x}_i,\mathbf{x}_0)}\left[\log\frac{p(\mathbf{x}_{i-1}|\mathbf{x}_i,\mathbf{x}_0)}{p(\mathbf{x}_{i-1}|\mathbf{x}_i)}\right]\right]\right]\\ &+\mathbb{E}_{p(\mathbf{x}_i)}\left[\mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_i)}\left[\mathbb{E}_{p(\mathbf{x}_{i-1}|\mathbf{x}_i,\mathbf{x}_0)}\left[\log\frac{p(\mathbf{x}_{i-1}|\mathbf{x}_i)}{p_\phi(\mathbf{x}_{i-1}|\mathbf{x}_i)}\right]\right]\right].\end{aligned}$$

Since the second logarithmic term does not depend on $\mathbf{x}_0$, by the law of total probability

$$\mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_i)}\left[\mathbb{E}_{p(\mathbf{x}_{i-1}|\mathbf{x}_i,\mathbf{x}_0)}\left[\log\frac{p(\mathbf{x}_{i-1}|\mathbf{x}_i)}{p_\phi(\mathbf{x}_{i-1}|\mathbf{x}_i)}\right]\right] = \mathbb{E}_{p(\mathbf{x}_{i-1}|\mathbf{x}_i)}\left[\log\frac{p(\mathbf{x}_{i-1}|\mathbf{x}_i)}{p_\phi(\mathbf{x}_{i-1}|\mathbf{x}_i)}\right].$$

Similarly, the first term is the KL divergence

$$\mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_i)}\left[\mathcal{D}_{\mathrm{KL}}(p(\mathbf{x}_{i-1}|\mathbf{x}_i,\mathbf{x}_0)\|p(\mathbf{x}_{i-1}|\mathbf{x}_i))\right].$$

Putting it all together, we obtain the decomposition:

$$\begin{aligned}&\mathbb{E}_{p(\mathbf{x}_0,\mathbf{x}_i)}\left[\mathcal{D}_{\mathrm{KL}}(p(\mathbf{x}_{i-1}|\mathbf{x}_i,\mathbf{x}_0)\|p_\phi(\mathbf{x}_{i-1}|\mathbf{x}_i))\right]\\ =&\mathbb{E}_{p(\mathbf{x}_i)}\left[\mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_i)}\left[\mathcal{D}_{\mathrm{KL}}(p(\mathbf{x}_{i-1}|\mathbf{x}_i,\mathbf{x}_0)\|p(\mathbf{x}_{i-1}|\mathbf{x}_i))\right]\right]\\ &+\mathbb{E}_{p(\mathbf{x}_i)}\left[\mathcal{D}_{\mathrm{KL}}(p(\mathbf{x}_{i-1}|\mathbf{x}_i)\|p_\phi(\mathbf{x}_{i-1}|\mathbf{x}_i))\right].\end{aligned}$$

**Proof of Optimality.** To prove:

$$p^*(\mathbf{x}_{i-1}|\mathbf{x}_i) = p(\mathbf{x}_{i-1}|\mathbf{x}_i) = \mathbb{E}_{p(\mathbf{x}|\mathbf{x}_i)}\left[p(\mathbf{x}_{i-1}|\mathbf{x}_i,\mathbf{x})\right], \quad \mathbf{x}_i \sim p_i.$$

The first identity follows from the fact that the KL divergence $\mathcal{D}_{\mathrm{KL}}(p\|p_\phi)$ is minimized when $p^* = p$, assuming the parameterization is sufficiently expressive. The second identity follows directly from the law of total probability. ■

Conditioning to obtain a tractable objective, forms the foundation of DDPMs.

- **DDPM Sampling:**
  - $\epsilon_{\phi x}(x_i, i)$ $i = 1, 2, \cdots, L$ ← after training
  - $X_L \sim P_{piror} = N(0, I)$ : sampling procedure

$$X_L \xrightarrow{P_{\phi x}(X_{L-1}|X_L)} X_{L-1} \xrightarrow{P_{\phi x}(X_{L-2}|X_{L-1})} X_{L-2} \cdots \xrightarrow{P_{\phi x}(X_0|X_1)} X_0$$

$$x_{i-1} \leftarrow \underbrace{\frac{1}{\alpha_i}\left(x_i - \frac{1-\alpha_i^2}{\sqrt{1-\bar{\alpha}_i^2}}\epsilon_{\phi x}(x_i, i)\right)}_{M_{\phi x}(x_i, i)} + \sigma(i)\cdot\epsilon_i, \quad \epsilon_i \sim N(0, I)$$

  - Interpretation of DDPM's sampling:

> $x_i \leftarrow$ interpolation between $x_i$ and $x_{\phi x}(x_i, i)$ + $\sigma(i)\cdot\epsilon_i$
>
> $\Leftrightarrow x_i = A\cdot x_i + B\cdot x_{\phi x}(x_i, i) + \sigma(i)\cdot\epsilon_i$

Proof: By $x_i = \bar{\alpha}_i \cdot x_\phi(x_i, i) + \sqrt{1-\bar{\alpha}_i^2}\,\epsilon_\phi(x_i, i)$

$$\Rightarrow X_{\phi x}(x_i, i) = \frac{x_i - \sqrt{1-\bar{\alpha}_i^2}\cdot\epsilon_{\phi x}(x_i, i)}{\bar{\alpha}_i}$$

Rearranging to isolate $\epsilon_{\phi x}$ :

$$\epsilon_{\phi x}(x_i, i) = \frac{x_i - \bar{\alpha}_i \cdot x_{\phi x}(x_i, i)}{\sqrt{1-\bar{\alpha}_i^2}}$$

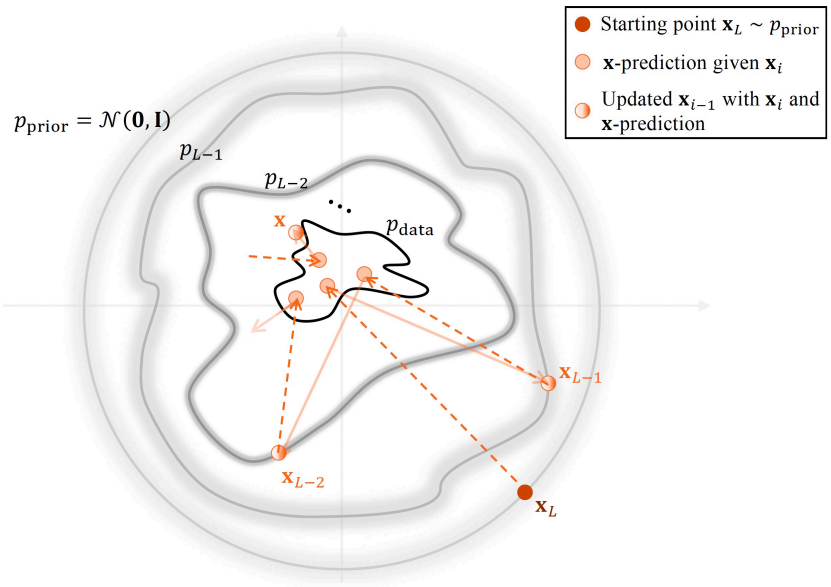
$$\Rightarrow M_{\phi x}(x_i, i) = \underbrace{\left(\frac{1}{\alpha_i} - \frac{1-\alpha_i^2}{\alpha_i(1-\bar{\alpha}_i^2)}\right)}_{A}\cdot x_i + \underbrace{\left(\frac{(1-\alpha_i^2)\cdot\bar{\alpha}_i}{\alpha_i(1-\bar{\alpha}_i^2)}\right)}_{B}\cdot x_{\phi x}(x_i, i)$$

■

DDPM can be viewed as an iterative denoising process:

(1): Estimating clean data $x_{\phi^*}(x_i, i)$ from current noisy input $x_i$.

(2): Sampling less noisy latent $x_{i-1}$ via the update rule using this clean estimate.




- Remarks of sampling via DDPM:

(1): Blurry predictions at high noise levels

$$x^*(x_i, i) = \mathbb{E}[x_0 \mid x_i], \quad x_i \sim p_i$$

the model predict the expected clean data given $x_i$

(2): Early steps set the global structure
Later steps add fine detail

(3): slow sampling:
$L = 1000$ steps in practice

(4) In practice, $p_\phi(x_{i-1} \mid x_i)$ modeled as Gaussian

small $\beta_i$ → true reverse distribution closes to Gaussian

large $\beta_i$ → induce strong non-Gaussian.

References:

[1] https://lilianweng.github.io/posts/2018-08-12-vae/