Denoising Diffusion Probabilistic Models Notes (DDPM)

Jayin Khanna
Shiv Nadar Institute of Eminence

December 24, 2025

Data and Goal

Given data

$$\mathcal{D} = \{x_1, x_2, \ldots, x_n\}, \quad x_i \sim p_X \text{ i.i.d.}$$

Goal: Learn a procedure to sample from the data distribution p_X.

DDPMs as Latent Variable Generative Models

Denoising Diffusion Probabilistic Models (DDPMs) are latent variable generative models.

They can be viewed as a structured special case of *Hierarchical Variational Autoencoders (HVAE)*.

VAE Reminder

In a standard VAE:

$$x \xrightarrow{\text{encoder}} z \xrightarrow{\text{decoder}} x$$

- Encoder: $q_\phi(z \mid x)$ - Decoder: $p_\theta(x \mid z)$

Hierarchical VAE View

In an HVAE, there are multiple latent spaces:

$$x \to z_1 \to z_2 \to \cdots \to z_T$$

(encoding)

$$z_T \to z_{T-1} \to \cdots \to z_1 \to x$$

(decoding)

Why an HVAE?

DDPMs require multiple latent variables to model gradual corruption and gradual denoising.

DDPM as an HVAE: Key Properties

1. Multiple Latent Spaces There exist latent variables

$$z_1, z_2, \ldots, z_T$$

2. Latent Dimensionality The dimensionality of each latent space equals that of the data:

$$\dim(z_t) = \dim(x), \quad \forall t$$

(Possible reason) Compression leads to loss of information; diffusion preserves dimensionality.

3. Fixed Encoding Procedure The encoding process is fixed (not learnable), unlike a VAE.

$$\text{VAE: } q_\phi(z \mid x) \quad \text{learned}$$

$$\text{DDPM: } q(z \mid x) \quad \text{fixed}$$

Only the decoding (denoising) process is learned.

DDPM Formulation

Notation

- Data variable: $x_0 \sim p_X$ - Latent variables: $x_1, x_2, \ldots, x_T$

Forward (Encoding) Process

$$x_0 \to x_1 \to x_2 \to \cdots \to x_T$$

Each x_t is a latent version of the original data x_0.

Dimensionality

$$\dim(x_t) = \dim(x_0) = p, \quad \forall t$$

The number of steps T is a hyperparameter.

Forward Process Definition

The forward diffusion process is defined as:

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

where

$$\alpha_t \in (0, 1]$$

Expanded Form

$$\begin{aligned} x_1 &= \sqrt{\alpha_1} x_0 + \sqrt{1 - \alpha_1} \varepsilon_1 \\ x_2 &= \sqrt{\alpha_2} x_1 + \sqrt{1 - \alpha_2} \varepsilon_2 \\ &\vdots \\ x_T &= \sqrt{\alpha_T} x_{T-1} + \sqrt{1 - \alpha_T} \varepsilon_T \end{aligned}$$

Intuition: The data is gradually destroyed by Gaussian noise until it becomes nearly isotropic noise.

Markov Structure of the Forward Process

From the forward definition

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

the process forms a first-order Markov chain:

$$x_t \perp \{x_{t-2}, \ldots, x_0\} \mid x_{t-1}.$$

Conditional Distribution of the Forward Process

The conditional distribution of each step is Gaussian:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{\alpha_t}\, x_{t-1}, (1 - \alpha_t) I\right).$$

Intuition: Each step shrinks the previous signal and injects fresh Gaussian noise.

Stationary Distribution

As $t \to T$, repeated noise injection causes the distribution of x_t to converge to:

$$q(x_T) \approx \mathcal{N}(0, I).$$

This serves as the prior distribution for the reverse process.

Joint Distribution of the Model

The generative (reverse) model is defined via a decoding distribution:

$$p_\theta(x_0, x_1, \ldots, x_T) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t),$$

where

$$p(x_T) = \mathcal{N}(0, I),$$

and

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

Both μ_θ and Σ_θ are learned via optimization.

Learning Objective

The model parameters θ are learned by maximizing the log-likelihood:

$$\log p_\theta(x_0).$$

Since the marginal likelihood is intractable, we instead optimize a variational lower bound.

ELBO in a VAE

Recall the VAE objective:

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

Introducing a variational distribution $q(z \mid x)$,

$$\log p_\theta(x) \geq \mathbb{E}_{q(z|x)} \left[\log \frac{p_\theta(x, z)}{q(z \mid x)}\right] \triangleq \mathcal{L}_{\text{ELBO}}.$$

ELBO for DDPM

For DDPMs, define:

$$x_{1:T} = (x_1, x_2, \ldots, x_T), \quad x_{0:T} = (x_0, x_1, \ldots, x_T).$$

The ELBO becomes:

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

Factorizations

From the generative model:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t). \tag{2}$$

From the forward process:

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}). \tag{3}$$

The forward encoding is a first-order Markov chain.

Substituting into the ELBO

From (1), (2), and (3):

$$\mathcal{L}(q) = \mathbb{E}\left[\log \frac{p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t)}{\prod_{t=1}^{T} q(x_t \mid x_{t-1})}\right].$$

Conditioning on the Original Data

We introduce conditioning on x_0 to simplify the ELBO.

Note that:

$$q(x_{t-1} \mid x_t, x_0)$$

denotes the distribution of x_{t-1} given that the forward process started from x_0 and reached x_t at step t.

> *Interpretation:* Since the forward process is known and fixed, conditioning on x_0 provides additional information that makes posterior inference tractable.

Why Condition on x_0?

- Without conditioning on x_0, the posterior $q(x_{t-1} \mid x_t)$ is intractable.
- Conditioning on x_0 anchors the trajectory to a known origin.
- This yields a computable Gaussian posterior.

Rewriting the ELBO

Starting from:

$$\mathcal{L}(q) = \mathbb{E}\left[\log \frac{p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t)}{\prod_{t=1}^{T} q(x_t \mid x_{t-1})}\right],$$

rewrite as:

$$= \log \frac{p(x_T)p_\theta(x_0 \mid x_1)\prod_{t=2}^{T} p_\theta(x_{t-1} \mid x_t)}{q(x_1 \mid x_0)\prod_{t=2}^{T} q(x_t \mid x_{t-1})}. \tag{5}$$

Separating terms:

$$= \log \frac{p_\theta(x_0 \mid x_1)}{q(x_1 \mid x_0)} + \log \frac{p(x_T)}{q(x_T \mid x_0)} + \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1} \mid x_t)}{q(x_{t-1} \mid x_t, x_0)}.$$

Using Bayes' Rule

Observe that:

$$q(x_{t-1} \mid x_t, x_0) = \frac{q(x_t \mid x_{t-1}) q(x_{t-1} \mid x_0)}{q(x_t \mid x_0)}.$$

Reason: We use Bayes' rule to express all quantities in terms of the known forward process.

Expectation Structure

Using the law of total expectation:

$$\mathbb{E}_{q(x_{1:T}|x_0)} = \mathbb{E}_{q(x_t|x_0)} \mathbb{E}_{q(x_{t-1}|x_t, x_0)}.$$

Final ELBO Decomposition

The ELBO can now be written as:

$$\mathcal{L}(q) = \mathbb{E}_{q(x_1|x_0)} \left[\log p_\theta(x_0 \mid x_1)\right] - \mathrm{D}_{\mathrm{KL}} \left(q(x_T \mid x_0) \,\|\, p(x_T)\right)$$

$$- \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)} \left[\mathrm{D}_{\mathrm{KL}} \left(q(x_{t-1} \mid x_t, x_0) \,\|\, p_\theta(x_{t-1} \mid x_t)\right)\right].$$

Interpretation of Terms

The DDPM objective consists of:

1. **Reconstruction term:**
$$\mathbb{E}_{q(x_1|x_0)} \left[\log p_\theta(x_0 \mid x_1)\right]$$

2. **Prior matching term:**
$$\mathrm{D}_{\mathrm{KL}} \left(q(x_T \mid x_0) \,\|\, p(x_T)\right)$$

3. **Consistency (denoising) terms:**
$$\sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)} \left[\mathrm{D}_{\mathrm{KL}} \left(q(x_{t-1} \mid x_t, x_0) \,\|\, p_\theta(x_{t-1} \mid x_t)\right)\right]$$

Key point: The true posterior is known; the model learns to match it using a parameterized reverse process.

Closed-Form of the Forward Process

Recall the forward process:

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

By recursively substituting earlier steps, we obtain a closed-form expression for x_t in terms of x_0.

Cumulative Product Notation

Define:

$$\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s.$$

Marginal Distribution of $x_t \mid x_0$

Using repeated substitution:

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

Thus,

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t} x_0, (1-\bar{\alpha}_t) I\right).$$

Key idea: Noise accumulates additively, while signal decays multiplicatively.

Posterior Distribution $q(x_{t-1} \mid x_t, x_0)$

Both:

$$q(x_{t-1} \mid x_0), \quad q(x_t \mid x_{t-1})$$

are Gaussian.

Therefore, the posterior

$$q(x_{t-1} \mid x_t, x_0)$$

is also Gaussian.

Bayesian Derivation

Using Bayes' rule:

$$q(x_{t-1} \mid x_t, x_0) \propto q(x_t \mid x_{t-1}) q(x_{t-1} \mid x_0).$$

Substituting known forms:

$$q(x_{t-1} \mid x_t, x_0) = \mathcal{N}\left(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I\right).$$

Posterior Variance

The variance is given by:

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

Posterior Mean

The posterior mean is:

$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)}{1 - \bar{\alpha}_t}x_0.$$

Interpretation: The posterior mean interpolates between the noisy observation x_t and the original data x_0.

Connection to the Reverse Model

The true reverse distribution:

$$q(x_{t-1} \mid x_t, x_0)$$

is known and Gaussian.

The learned reverse model:

$$p_\theta(x_{t-1} \mid x_t)$$

is trained to approximate this distribution.

This leads to a KL divergence term in the ELBO:

$$\mathrm{D}_{\mathrm{KL}}\left(q(x_{t-1} \mid x_t, x_0) \parallel p_\theta(x_{t-1} \mid x_t)\right).$$

Minimizing this KL encourages the learned reverse process to match the true denoising posterior.

Parameterization of the Reverse Process

Recall the learned reverse distribution:

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

In practice, the variance is often fixed:

$$\Sigma_\theta(x_t, t) = \beta_t I,$$

and only the mean is learned.

Predicting the Mean via Noise Estimation

From the closed-form forward process:

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

we can solve for x_0:

$$x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(x_t - \sqrt{1 - \bar{\alpha}_t}\, \varepsilon \right).$$

Modeling the Noise

Instead of predicting x_0 directly, the model predicts the noise:

$$\varepsilon_\theta(x_t, t) \approx \varepsilon.$$

Reason: Noise has a simpler, isotropic structure compared to data.

Mean in Terms of Predicted Noise

Substituting x_0 into the posterior mean:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \varepsilon_\theta(x_t, t) \right).$$

This is the parameterization used in DDPMs.

Simplified Training Objective

The KL divergence between two Gaussians reduces to a squared error.

Thus, the ELBO simplifies to:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t, x_0, \varepsilon} \left[\| \varepsilon - \varepsilon_\theta(x_t, t) \|^2 \right],$$

where

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\, \varepsilon.$$

Training Procedure

1. Sample $x_0 \sim p_X$
2. Sample $t \sim \text{Uniform}\{1, \ldots, T\}$
3. Sample $\varepsilon \sim \mathcal{N}(0, I)$
4. Construct x_t

5. Minimize $\|\varepsilon - \varepsilon_\theta(x_t, t)\|^2$

Key idea: Training reduces to denoising Gaussian noise at arbitrary time steps.

Sampling (Generation)

Generation starts from pure noise:

$$x_T \sim \mathcal{N}(0, I).$$

For $t = T, T-1, \ldots, 1$:

$$x_{t-1} = \mu_\theta(x_t, t) + \sqrt{\beta_t} z, \quad z \sim \mathcal{N}(0, I),$$

except at $t = 1$, where no noise is added.

The final sample x_0 is a generated data point.

Interpretation: Sampling reverses the diffusion process step by step.