

Study Material - Youtube

Document Information

- **Generated:** 2025-08-01 22:27:38
- **Source:** <https://youtube.com/watch?v=P8AiIW0Gg0s>
- **Platform:** Youtube
- **Word Count:** 2,033 words
- **Estimated Reading Time:** ~10 minutes
- **Number of Chapters:** 6
- **Transcript Available:** Yes (analyzed from video content)

Table of Contents

1. DDPM: Formulation and Core Concepts
 2. Mathematical Analysis of the Forward Process (Encoding)
 3. Mathematical Analysis of the Reverse Process (Decoding)
 4. Training Objective: The Evidence Lower Bound (ELBO)
 5. Self-Assessment for This Video
 6. Key Takeaways from This Video
-

Video Overview

This video lecture, titled “DDPM: Formulation,” is part of a broader course on the “Mathematical Foundations of Generative AI.” The instructor, Prof. Prathosh A P, provides a detailed mathematical formulation of Denoising Diffusion Probabilistic Models (DDPMs). The lecture begins by contrasting DDPMs with Variational Autoencoders (VAEs), highlighting the fundamental difference in their learning mechanisms. It then establishes the notational conventions used in DDPM literature before delving into the two core components of the model: the **forward (encoding) process** and the **reverse (decoding) process**. The lecture culminates in setting up the Evidence Lower Bound (ELBO) as the optimization objective for training DDPMs, laying the groundwork for subsequent derivations.

Learning Objectives

Upon completing this lecture, a student should be able to: * Understand the core conceptual difference between VAEs and DDPMs, specifically that DDPMs have a fixed, non-learnable encoding process. * Recognize and use the standard notation for data and latent variables in the context of DDPMs. * Describe the forward (diffusion) process as a Markov chain that progressively adds Gaussian noise to data. * Write down and explain the mathematical equation governing the forward process transitions. * Describe the reverse (denoising) process as a learned Markov chain designed to undo the diffusion. * Formulate the joint probability distribution of the DDPM model. * Set up the Evidence Lower Bound (ELBO) objective function for training a DDPM.

Prerequisites

To fully grasp the concepts in this video, students should have a solid understanding of: * **Probability Theory:** Concepts like conditional probability, joint distributions, Gaussian distributions, and the chain rule of probability. * **Calculus and Linear Algebra:** Basic calculus and vector/matrix operations. * **Machine Learning Fundamentals:** Familiarity with generative models. * **Variational Autoencoders (VAEs):** A strong understanding of VAE architecture, including the encoder-decoder structure and the Evidence Lower Bound (ELBO), is essential, as the lecture frequently draws comparisons.

Key Concepts

- Denoising Diffusion Probabilistic Models (DDPMs)
 - Forward Process (Diffusion/Encoding)
 - Reverse Process (Denoising/Decoding)
 - Markov Chains
 - Latent Variables
 - Evidence Lower Bound (ELBO)
-

DDPM: Formulation and Core Concepts

1. Foundational Distinction: DDPM vs. VAE

(0:10) The lecture begins by establishing a critical distinction between Denoising Diffusion Probabilistic Models (DDPMs) and Variational Autoencoders (VAEs). This difference is central to understanding the DDPM architecture.

- **Variational Autoencoder (VAE):** In a VAE, **both the encoding and decoding processes are learned**. The encoder, typically denoted as $q_\phi(z|x)$, maps data x to a latent representation z and has learnable parameters ϕ . The decoder, $p_\theta(x|z)$, maps the latent representation back to the data space and has learnable parameters θ .
- **Denoising Diffusion Probabilistic Model (DDPM):** In a DDPM, **only the decoding process is learned**. The encoding process is a fixed, non-learnable procedure.

Key Insight: The primary innovation in DDPMs is the introduction of a fixed, predefined encoding (forward) process. This simplifies the learning problem to only discovering the corresponding decoding (reverse) process.

2. The Two Processes of DDPMs

(0:21) A DDPM is characterized by two opposing processes: a forward process that corrupts data with noise and a reverse process that learns to remove it.

1. **Forward Process (Encoding / Diffusion):** This is a fixed, non-learnable Markov process that gradually adds Gaussian noise to an input data point (e.g., an image) over a sequence of T timesteps. It starts with clean data and ends with something that is indistinguishable from pure noise.
2. **Reverse Process (Decoding / Denoising):** This is a learnable Markov process that aims to reverse the forward process. It starts with pure noise and iteratively removes noise at each timestep to generate a clean data sample. The “denoising” aspect of the model’s name refers to this learned reverse process.

The relationship between these processes can be visualized as follows:

```
sequenceDiagram
    participant D as Data (x)
    participant L1 as Latent (x)
    participant L2 as Latent (x)
    participant LT as Latent (xT) / Noise

    box rgb(255, 240, 240) Forward Process (Fixed)
        D->>L1: Add Noise
        L1->>L2: Add More Noise
        L2->>LT: ...
    end

    box rgb(240, 240, 255) Reverse Process (Learned)
```

```

    LT->>L2: Denoise
    L2->>L1: Denoise More
    L1->>D: ...
end

```

Figure 1: A conceptual diagram illustrating the fixed forward (noising) process and the learned reverse (denoising) process in a DDPM.

3. Notational Conventions in DDPMs

(0:57) To align with the standard DDPM literature, the instructor introduces a specific set of notations that differ from those typically used for VAEs.

- **Data Variable:** The original, clean data point is denoted by \mathbf{x}_0 . This corresponds to what might be called x in other contexts.
- **Latent Variables:** The sequence of noisy versions of the data is represented by $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$. These are the latent variables of the model.
 - x_t is the data after t steps of adding noise.
 - This sequence of latent variables replaces the single latent variable z found in a basic VAE.
- **Dimensionality:** A key property of DDPMs is that the dimensionality of all latent variables is the same as the original data:

$$\dim(x_t) = \dim(x_0) \quad \forall t \in \{1, \dots, T\}$$

- **Timesteps (T):** The total number of diffusion steps, T , is a hyperparameter. It is typically a large number (e.g., 1000).

Warning: It is crucial not to confuse the indexed variables x_1, x_2, \dots, x_T with different samples from a dataset. They represent a single data point, x_0 , at different stages of the noising process.

Mathematical Analysis of the Forward Process (Encoding)

(5:57) The forward process, also known as the diffusion process, is a fixed Markov chain that progressively adds Gaussian noise to the data.

The Forward Process as a Markov Chain

The process is defined as a sequence of latent variables x_1, \dots, x_T generated from the initial data x_0 . The generation of x_t depends only on the previous state x_{t-1} , which is the defining property of a **first-order Markov chain**.

The entire forward process is defined by the joint distribution of the latent variables conditioned on the starting data, which can be factored due to the Markov property:

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

The Forward Step Equation

(9:04) Each transition in the forward process, from x_{t-1} to x_t , is defined by the following equation:

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

Note: The video uses ϵ_t in the equation for x_t , but later clarifies the process is iterative. For clarity, we use ϵ_{t-1} to generate x_t from x_{t-1} .

Intuitive Breakdown: * \mathbf{x}_{t-1} : The (less noisy) data at the previous timestep. * \mathbf{x}_t : The (more noisy) data at the current timestep. * α_t : A hyperparameter from a predefined “variance schedule” where $0 < \alpha_t < 1$. It

controls how much of the previous state is preserved. * ϵ_{t-1} : A random noise vector sampled from a standard Gaussian distribution. * $\sqrt{\alpha_t}x_{t-1}$: This term scales down the previous state. * $\sqrt{1-\alpha_t}\epsilon_{t-1}$: This term adds new noise, with its variance controlled by α_t . The coefficients $\sqrt{\alpha_t}$ and $\sqrt{1-\alpha_t}$ are chosen such that the overall variance of the process is well-behaved.

Conditional Distribution of the Forward Process

(22:42) From the forward step equation, we can define the conditional probability distribution for each transition. Since x_t is a linear transformation of a Gaussian variable (x_{t-1} , treated as a constant here) plus another Gaussian variable (ϵ_{t-1}), the resulting distribution is also Gaussian.

$$q(x_t|x_{t-1}) \triangleq \mathcal{N}(x_t; \mu_t, \sigma_t^2 I)$$

where: * **Mean:** $\mu_t = \sqrt{\alpha_t}x_{t-1}$ * **Variance:** $\sigma_t^2 I = (1 - \alpha_t)I$

So, the full conditional distribution is:

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

This process is repeated for T steps. The variance schedule (the values of α_t) is chosen such that after T steps, the final latent variable x_T is approximately distributed as a standard isotropic Gaussian, i.e., $q(x_T|x_0) \approx \mathcal{N}(0, I)$.

Mathematical Analysis of the Reverse Process (Decoding)

(25:33) The reverse process is the core of the generative model. It learns to reverse the diffusion, starting from a random noise vector $x_T \sim \mathcal{N}(0, I)$ and producing a clean data sample x_0 .

The Reverse Process as a Learned Markov Chain

The reverse process is also modeled as a Markov chain, running from $t = T$ down to $t = 1$. The goal is to learn the transition probabilities $p_\theta(x_{t-1}|x_t)$.

The joint distribution over the latent variables, parameterized by θ , is given by:

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

Breakdown of Terms: * $p(x_T) = \mathcal{N}(x_T; 0, I)$: The prior distribution, which is a standard Gaussian. * $p_\theta(x_{t-1}|x_t)$: The reverse transition probability. This is what the model learns. It represents the probability of transitioning from a more noisy state x_t to a slightly less noisy state x_{t-1} .

Gaussian Transition Assumption

(27:28) The reverse transitions are also parameterized as Gaussian distributions:

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

- $\mu_\theta(x_t, t)$: The mean of the distribution, predicted by a neural network.
- $\Sigma_\theta(x_t, t)$: The variance of the distribution, also predicted by a neural network.

The neural network takes the current noisy image x_t and the timestep t as input and outputs the parameters for the Gaussian distribution of the previous, less noisy image x_{t-1} . The parameters of this network, θ , are what we optimize during training.

Training Objective: The Evidence Lower Bound (ELBO)

(37:11) To learn the parameters θ of the reverse process, we use the same principle as in VAEs: maximizing the Evidence Lower Bound (ELBO) on the log-likelihood of the data.

ELBO for DDPMs

(40:17) For a DDPM with a sequence of latent variables $x_{1:T}$, the ELBO is formulated as the expectation over the *fixed* forward process q of the log-ratio between the *learned* reverse process p_θ and the forward process q .

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

Key Components: * $\mathbb{E}_{q(x_{1:T}|x_0)}[\cdot]$: The expectation is taken with respect to the joint distribution of all latent variables generated by the fixed forward process. * $p_\theta(x_{0:T})$: The joint distribution of the model (the reverse process). This is the term that contains the learnable parameters θ . * $q(x_{1:T}|x_0)$: The joint distribution of the forward process, conditioned on the initial data. This term is fixed and has no learnable parameters.

The goal of training is to find the parameters θ that maximize this objective function. The subsequent steps in the full derivation of DDPMs involve simplifying this ELBO into a more tractable form, which typically decomposes into a sum of KL-divergence terms for each timestep.

Self-Assessment for This Video

1. Conceptual Questions:

- What is the fundamental difference between how a VAE and a DDPM handle the encoding process?
- Explain the terms “forward process” and “reverse process” in your own words. Which one is learned?
- Why is the forward process in a DDPM considered a Markov chain?
- What is the role of the hyperparameter T ? What happens to the data x_t as $t \rightarrow T$?

2. Mathematical Questions:

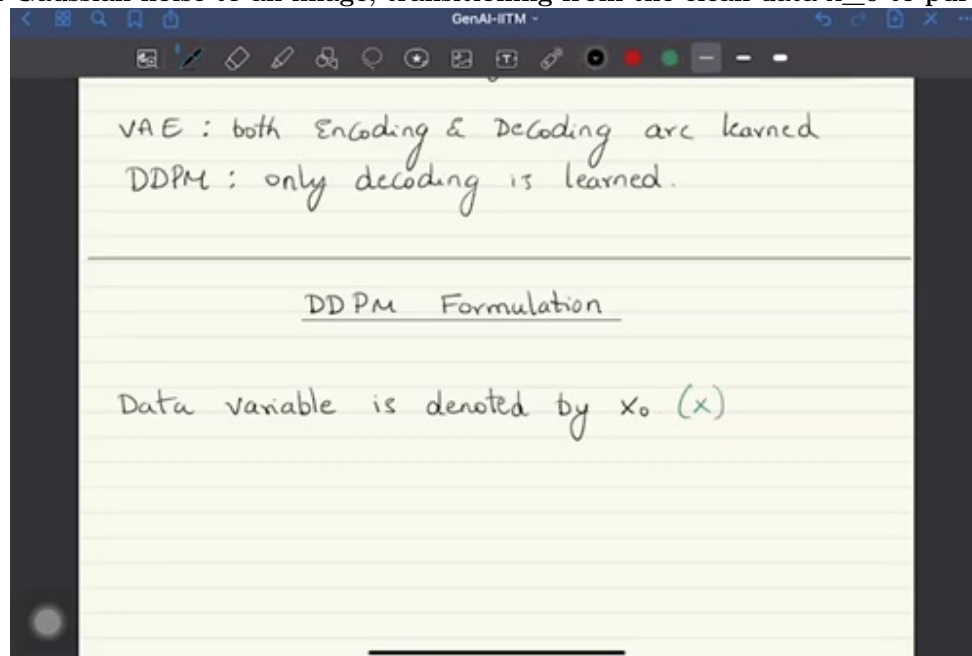
- Write down the equation for a single step in the forward diffusion process. Explain what each variable and parameter represents.
- Given x_{t-1} , what is the distribution of $q(x_t|x_{t-1})$? Specify its mean and variance.
- Write the mathematical form of the joint distribution $p_\theta(x_{0:T})$ for the reverse process. What parts of this expression are learnable?
- What is the ELBO for a DDPM? Explain why the optimization is performed with respect to θ but not with respect to any parameters of q .

Key Takeaways from This Video

- **DDPMs simplify the generative modeling problem by fixing the encoding (forward) process.** The model only needs to learn the reverse (decoding) process.
- The **forward process** is a Markov chain that systematically adds Gaussian noise to data over T steps, transforming it into pure noise.
- The **reverse process** is a learned Markov chain that aims to reverse the noising, step-by-step, to generate data from noise.
- The transitions in both the forward and reverse processes are modeled as **Gaussian distributions**.
- The parameters of the reverse process’s Gaussian transitions are predicted by a neural network, which is trained by **maximizing the Evidence Lower Bound (ELBO)**.

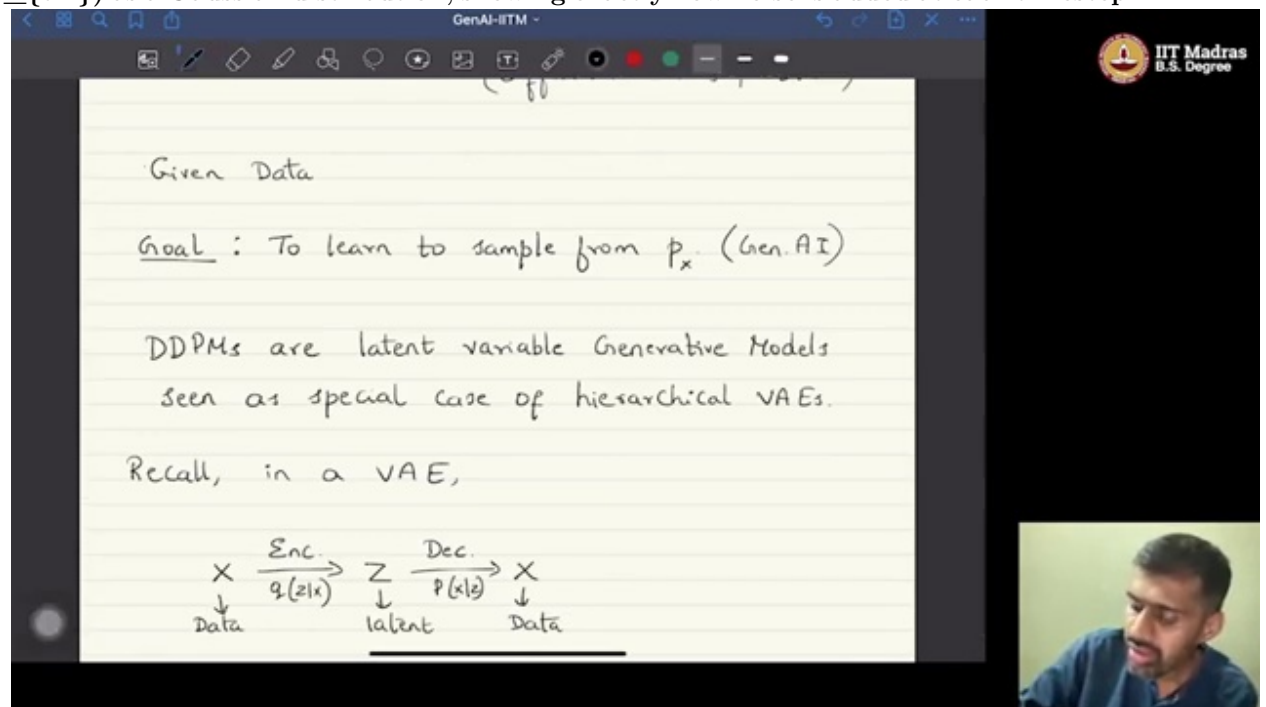
Visual References

A diagram illustrating the forward process as a Markov chain. It visually represents the sequential addition of Gaussian noise to an image, transitioning from the clean data x_0 to pure noise



x_T . (at 01:45):

The core mathematical equation for a single step in the forward (diffusion) process. It defines $q(x_t | x_{t-1})$ as a Gaussian distribution, showing exactly how noise is added at each timestep.



(at 02:58):

The equation for the parameterized reverse process, $p_{\theta}(x_{t-1}|x_t)$. This is a key slide as it shows that the reverse transition is a Gaussian whose mean and variance are predicted by a neu-

GenAI-IITM


DDPM Formulation

Notations :

Data variable is denoted by $x_0 (x)$

Latent space : $x_1, x_2, \dots, x_T (z_1, z_2, \dots, z_T)$

(not be to confused with previous notations where x_1, x_2, \dots were data points) in the DI



ral network. (at 05:20):

The formulation of the Evidence Lower Bound (ELBO) as the training objective for the DDPM. This equation is the foundation for the model's entire training procedure. (at 06:40):

GenAI-IITM


where x_1, x_2, \dots were data points)

In the DDPM literature, x_0 : data point

x_1, x_2, \dots, x_T : latent spaces / vectors.

Encoding | Forward process :

$x_0 \rightarrow x_1 \rightarrow x_2 \rightarrow \dots \rightarrow x_T$



IIT Madras
B.S. Degree