Study Material - Youtube

Document Information

- **Generated:** 2025-08-26 06:42:21
- **Source:** https://www.youtube.com/watch?v=P8AiIW0Gg0s
- **Platform:** Youtube
- **Word Count:** 1,968 words
- **Estimated Reading Time:** ~9 minutes
- **Number of Chapters:** 3
- **Transcript Available:** Yes (analyzed from video content)

Table of Contents

Video Overview

This lecture, titled "DDPM: Formulation," is part of the "Mathematical Foundations of Generative AI" series. 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), establishing that DDPMs have a fixed encoding process and a learned decoding process. The core of the lecture is dedicated to defining the two key components of a DDPM: the **forward (encoding/noising) process** and the **reverse (decoding/denoising) process**. The instructor carefully introduces the notation and mathematical equations that govern these processes, setting the stage for understanding how DDPMs are trained using the Evidence Lower Bound (ELBO) optimization, which will be covered in subsequent material.

Learning Objectives

Upon completing this study material, students will be able to: - Understand the fundamental difference between VAEs and DDPMs in terms of what is learned. - Grasp the new notational conventions used in DDPM literature, specifically for data and latent variables. - Explain the concept of the forward (noising) process as a fixed, first-order Markov chain. - Write and interpret the mathematical equations for the forward process, including the role of the variance schedule α_t. - Explain the concept of the reverse (denoising) process as a learned Markov chain. - Write and interpret the mathematical formulation for the reverse process, identifying the learnable parameters. - Understand the structure of the joint probability distribution (the model) in a DDPM. - Formulate the Evidence Lower Bound (ELBO) for a DDPM.

Prerequisites

To fully understand the concepts in this lecture, students should have a solid foundation in: - **Probability Theory:** Conditional probability, Gaussian distributions, random variables, and expectation. - **Calculus:** Basic differentiation and integration. - **Linear Algebra:** Vectors and matrices. - **Machine Learning:** Familiarity with generative models, particularly Variational Autoencoders (VAEs), including the concept of encoders, decoders, latent space, and the ELBO.

Key Concepts Covered in This Video

- Denoising Diffusion Probabilistic Models (DDPMs)
- Forward Process (Encoding/Noising)
- Reverse Process (Decoding/Denoising)

- Markov Chains
- Gaussian Transitions
- Variance Schedule
- Evidence Lower Bound (ELBO) for DDPMs

Denoising Diffusion Probabilistic Models (DDPMs): Formulation

Foundational Distinction: DDPM vs. VAE

(00:11) The lecture begins by drawing a critical distinction between Variational Autoencoders (VAEs) and Denoising Diffusion Probabilistic Models (DDPMs). 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 to a latent space, and the decoder, $p_\theta(x|z)$, maps from the latent space back to the data space. The parameters ϕ and θ are learned jointly.
- **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 "denoising" in DDPM is synonymous with the **decoding** process. The model's primary task is to learn how to reverse a noising process, which is equivalent to decoding a clean signal from a noisy one.

The two main processes in a DDPM are: 1. **Encoding Process:** Also known as the **Forward Process**. This is a fixed, predefined procedure that gradually adds noise to the data. 2. **Decoding Process:** Also known as the **Reverse Process** or **Denoising Process**. This is the part of the model that is learned. It aims to reverse the forward process by gradually removing noise.

Notational Conventions for DDPMs

(01:05) To align with the standard DDPM literature, the instructor introduces a specific set of notations that differ from those typically used for VAEs. It is crucial to understand this change to follow the mathematical derivations.

Concept	VAE Notation	DDPM Notation	Explanation
Data Variable	x	x_0	The original, clean data point (e.g., an image) is considered the state at time $t = 0$.
Latent Variables	z	$x_1, x_2, \dots, x_T$	DDPMs use a sequence of latent variables. The variable x_t represents the data after t steps of adding noise. These are **not** different data samples but rather increasingly noisy versions of the same initial data point x_0.

This can be visualized as a sequence:

```
sequenceDiagram
    participant D as Data (x_0)
    participant L1 as Latent (x_1)
    participant L2 as Latent (x_2)
    participant LT as Latent (x_T)
    D->>L1: Add noise (step 1)
    L1->>L2: Add more noise (step 2)
    L2->>LT: ... (step T)
```

Figure 1: A conceptual diagram showing the sequential noising process in DDPMs, where each latent variable is a noisier version of the previous one.

The Forward Process (Encoding)

(05:57) The forward process is the fixed, non-learnable part of the DDPM that systematically corrupts the input data x_0 by adding Gaussian noise over T timesteps.

Intuitive Understanding

Imagine starting with a clear photograph (x_0). In each step of the forward process, you add a small amount of random, static-like noise. After one step, you get a slightly noisy image (x_1). After another step, you get an even noisier image (x_2), and so on. If you repeat this for a large number of steps (T), the final image (x_T) will be indistinguishable from pure random noise. The forward process defines exactly how this noise is added at each step.

Mathematical Formulation

(06:01) The forward process is defined as a **first-order Markov chain**. This means that the state at time t, denoted as x_t, depends only on the state at the previous time, $t-1$.

The transition probability, which defines how to get from x_{t-1} to x_t, is a Gaussian distribution:

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

Here, the instructor uses a slightly different but equivalent parameterization with $\alpha_t = 1 - \beta_t$. Let's stick to the notation used in the lecture for consistency. The reparameterized form is:

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

This leads to the conditional distribution:

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

Explanation of Terms: - x_t: The latent variable (noisy data) at timestep t. - x_{t-1}: The latent variable at the previous timestep. - α_t: A hyperparameter from a predefined **variance schedule**, where $0 < \alpha_t < 1$. It controls how much of the previous state's signal is preserved and how much noise is added. As t increases, α_t typically decreases, meaning more noise is added in later steps. - ϵ_{t-1}: A random noise vector sampled from a standard normal distribution $\mathcal{N}(0, I)$. - $\sqrt{\alpha_t}x_{t-1}$: This term scales down the previous state, preserving some of its structure. - $\sqrt{1-\alpha_t}\epsilon_{t-1}$: This term adds new noise, with its variance controlled by $1-\alpha_t$.

The entire forward process is the joint distribution of all latent variables given the initial data:

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

Key Properties of the Forward Process

1. **Fixed and Non-Learnable:** The parameters α_t are part of a predefined schedule (e.g., linear, cosine) and are not learned during training.
2. **Markovian:** The state x_t is conditionally independent of all past states $x_0, \dots, x_{t-2}$ given the immediate past state x_{t-1}.
3. **Dimensionality Preservation:** (07:16) The dimensionality of the latent variables is the same as the data space. If $x_0 \in \mathbb{R}^d$, then $x_t \in \mathbb{R}^d$ for all t.
4. **Stationary Distribution:** (36:14) For a sufficiently large T and a properly chosen schedule for α_t, the distribution of the final latent variable x_T converges to a standard isotropic Gaussian distribution:

$$q(x_T|x_0) \approx \mathcal{N}(0, I)$$

This means that after T steps, almost all information about the original data x_0 is lost, and what remains is pure noise.

The Reverse Process (Decoding)

(25:23) The reverse process is the core of the generative model. It is a learned process that aims to reverse the forward diffusion, starting from pure noise $x_T \sim \mathcal{N}(0, I)$ and gradually removing noise to generate a clean data sample x_0.

Mathematical Formulation

The reverse process is also modeled as a Markov chain, running from $t = T$ down to $t = 1$:

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

- $p(x_T) = \mathcal{N}(x_T; 0, I)$ is the prior, which is a standard Gaussian distribution. - $p_\theta(x_{t-1}|x_t)$ is the learned reverse transition probability. This is the "denoising" step.

The model assumes that these reverse transitions are also Gaussian:

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

Explanation of Terms: - $\mu_\theta(x_t, t)$: The mean of the distribution for the less-noisy state x_{t-1}, predicted by a neural network. The network takes the current noisy state x_t and the timestep t as input. - $\Sigma_\theta(x_t, t)$: The covariance of the distribution, also predicted by the neural network. In many DDPM papers, the covariance is kept fixed to simplify the model, and only the mean is learned. - θ: The learnable parameters of the neural network.

> **The Goal of Training:** The objective is to learn the parameters θ of the neural network such that the reverse process p_θ can accurately undo the fixed forward process q.

ELBO Optimization for DDPMs

(37:11) Since DDPMs are latent variable models, they are trained by maximizing the log-likelihood of the data, $\log p_\theta(x_0)$. As with VAEs, this is intractable, so we optimize its Evidence Lower Bound (ELBO).

The ELBO for a DDPM is:

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

Breakdown of the DDPM ELBO: - **Expectation:** The expectation is taken with respect to the **forward process** $q(x_{1:T}|x_0)$. This means during training, we take a data point x_0, generate a noisy sequence $x_1, \dots, x_T$

using the fixed forward process, and then evaluate the term inside the logarithm. - **Numerator $p_\theta(x_{0:T})$:** This is the joint probability of the **reverse process** (the model we are learning). - **Denominator $q(x_{1:T}|x_0)$:** This is the joint probability of the **forward process** (the fixed noising procedure).

Unlike in a VAE where the ELBO is optimized with respect to both encoder (ϕ) and decoder (θ) parameters, here the encoding distribution q is fixed. Therefore, we only need to optimize with respect to the decoder parameters θ.

Key Takeaways from This Video

- **DDPMs learn to denoise.** They are generative models that learn the reverse of a fixed, multi-step noising process.
- The **forward process** (encoding) is a fixed Markov chain that gradually adds Gaussian noise to data. It is not learned.
- The **reverse process** (decoding) is a learned Markov chain that aims to remove noise at each step, with its transition probabilities (mean and variance) parameterized by a neural network.
- **Notation is key:** The data is x_0, and the sequence of noisy latent variables is $x_1, \dots, x_T$.
- Training is performed by maximizing the **Evidence Lower Bound (ELBO)**, which involves optimizing the parameters of the reverse (decoding) process to best match the true data distribution.

Self-Assessment for This Video

1. **Conceptual Question:** What is the primary architectural difference between a VAE and a DDPM? Why is the encoding process in a DDPM referred to as "fixed"?
2. **Notation Check:** In the context of DDPMs, what do x_0, x_t, and x_T represent? How does this differ from the notation for data samples in a dataset like $\{x_1, x_2, \dots, x_N\}$?
3. **Forward Process Equation:**
   - Write down the equation for the conditional distribution $q(x_t|x_{t-1})$.
   - Explain the role of α_t and ϵ_{t-1} in this equation. What is a "variance schedule"?
4. **Reverse Process Formulation:**
   - Write the general form of the reverse transition probability $p_\theta(x_{t-1}|x_t)$.
   - What parts of this distribution are learned by the neural network?
5. **ELBO Formulation:** Write down the ELBO for a DDPM. Explain what distribution the expectation is taken over and which parameters are being optimized.