Study Material - Youtube

Document Information

- **Generated:** 2025-08-26 06:49:56
- **Source:** https://www.youtube.com/watch?v=wPx64rVy2c4
- **Platform:** Youtube
- **Word Count:** 2,149 words
- **Estimated Reading Time:** ~10 minutes
- **Number of Chapters:** 5
- **Transcript Available:** Yes (analyzed from video content)

Table of Contents

Video Overview

This video, "ELBO for DDPM: Part 2," is a continuation of the mathematical derivation of the Evidence Lower Bound (ELBO) for Denoising Diffusion Probabilistic Models (DDPMs). The lecturer focuses on rewriting the ELBO expression into a more interpretable and computationally tractable form. The core of the lecture involves algebraic manipulation using the properties of conditional expectation and Bayes' rule to express the ELBO in terms of KL divergence. The final objective function is shown to be composed of three distinct terms: a reconstruction term, a prior matching term, and a crucial consistency (or denoising matching) term. The lecture then delves into the detailed derivation of the posterior distribution required for the consistency term, highlighting the key simplifications that make training DDPMs feasible.

Learning Objectives

Upon completing this lecture, students will be able to:

- Understand how the ELBO for DDPMs can be decomposed into three meaningful terms.
- Apply the law of iterated expectations to simplify complex expectation expressions.
- Recognize and formulate KL divergence terms from the ELBO expression.
- Derive the closed-form expression for the posterior distribution $q(x_{t-1}|x_t, x_0)$ using Bayes' rule and properties of Gaussian distributions.
- Understand the recursive nature of the forward noising process and derive the distribution $q(x_t|x_0)$.
- Appreciate the final simplified form of the DDPM objective function and the role of each of its components.

Prerequisites

To fully grasp the concepts in this video, students should have a solid understanding of:

- **Probability Theory:** Conditional probability, Bayes' rule, expectation, variance, and properties of Gaussian distributions.
- **Information Theory:** Kullback-Leibler (KL) divergence and its properties.
- **Calculus:** Multivariate calculus, particularly differentiation.
- **Machine Learning:** Familiarity with Variational Autoencoders (VAEs) and the concept of the Evidence Lower Bound (ELBO).

- **Previous Lecture (W8L28):** Knowledge of the initial formulation of the ELBO for DDPMs is essential.

Key Concepts

- **ELBO Decomposition:** Breaking down the ELBO into a reconstruction term, a prior matching term, and a consistency term.
- **Law of Iterated Expectations:** Using conditional expectations to simplify complex probabilistic expressions.
- **Posterior Distribution $q(x_{t-1}|x_t, x_0)$:** The "true" reverse process distribution, which is tractable to compute.
- **Consistency (Denoising Matching) Term:** The core training objective that forces the learned reverse process to match the true reverse process.
- **Recursive Property of Forward Process:** The ability to sample any x_t directly from x_0 in a single step.

Deep Dive: Rewriting the ELBO for DDPMs

This section provides a detailed, step-by-step derivation of the simplified Evidence Lower Bound (ELBO) objective function for Denoising Diffusion Probabilistic Models (DDPMs), as presented in the lecture.

Rewriting the ELBO using Conditional Expectation

The lecture begins by recalling the ELBO expression derived in the previous session and aims to rewrite it into a more intuitive form. The initial expression for the ELBO, which we seek to maximize, is composed of three terms.

At (00:42), the instructor presents the ELBO expression:

$$L = \mathbb{E}_{q(x_1|x_0)}[\log p_\theta(x_0|x_1)] + \mathbb{E}_{q(x_T|x_0)}[\log \frac{p(x_T)}{q(x_T|x_0)}] + \sum_{t=2}^{T} \mathbb{E}_{q(x_{t-1},x_t|x_0)}[\log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}]$$

The goal is to simplify and interpret each of these terms. The key insight, introduced at (00:56), is to use the **property of conditional expectation** (also known as the law of total expectation or law of iterated expectations) to rewrite the expectation in the third term.

Property of Conditional Expectation

Intuition: This property allows us to compute a complex expectation by breaking it down into simpler, nested expectations. The overall expectation of a variable is the expected value of its conditional expectation.

Mathematical Formulation: For random variables A and B, the law states $\mathbb{E}[A] = \mathbb{E}_B[\mathbb{E}_{A|B}[A|B]]$. Applying this to our context:

$$\mathbb{E}_{q(x_{t-1},x_t|x_0)}[\cdot] = \mathbb{E}_{q(x_t|x_0)}\left[\mathbb{E}_{q(x_{t-1}|x_t,x_0)}[\cdot]\right]$$

This decomposition is visualized in the following flowchart:

```
flowchart TD
    A["Start with joint expectation<br/>E<sub>q(x<sub>t-1</sub>, x<sub>t</sub>|x<sub>0</sub>)</sub>[f(x
    B --> C["Outer Expectation<br/>E<sub>q(x<sub>t</sub>|x<sub>0</sub>)</sub>[...]"];
    C --> D["Inner Expectation<br/>E<sub>q(x<sub>t-1</sub>|x<sub>t</sub>, x<sub>0</sub>)</sub>[f(x<sub>t
```

Figure 1: Applying the law of iterated expectations to decompose the joint expectation.

Decomposing the ELBO into KL Divergence Terms

By applying this property and recognizing the definition of KL divergence, we can rewrite the ELBO.

1. **The First Term (Reconstruction Term):** This term remains as is. It measures how well the model can reconstruct the original data x_0 from the first noised sample x_1.

$$L_0 = \mathbb{E}_{q(x_1|x_0)}[\log p_\theta(x_0|x_1)]$$

2. **The Second Term (Prior Matching Term):** At (01:48), the instructor identifies this term as a KL divergence.

$$\mathbb{E}_{q(x_T|x_0)}[\log \frac{p(x_T)}{q(x_T|x_0)}] = -\mathbb{E}_{q(x_T|x_0)}[\log \frac{q(x_T|x_0)}{p(x_T)}] = -D_{KL}(q(x_T|x_0)||p(x_T))$$

   This term, which we can call L_T, forces the distribution of the fully noised data $q(x_T|x_0)$ to match a simple prior, typically a standard normal distribution $p(x_T) = \mathcal{N}(0, I)$. In practice, this term is often considered constant and ignored during optimization because the forward process is designed such that for a large T, $q(x_T|x_0)$ naturally approaches $\mathcal{N}(0, I)$.

3. **The Third Term (Consistency/Denoising Matching Term):** This is the most complex term. At (02:02), the instructor applies the conditional expectation property and rewrites it.

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

   Recognizing the inner expectation as a negative KL divergence, we get:

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

   This term is a sum of expected KL divergences. For each timestep t, it measures the "distance" between the true reverse step distribution $q(x_{t-1}|x_t, x_0)$ and the learned reverse step distribution $p_\theta(x_{t-1}|x_t)$. The goal of training is to make the learned denoising steps consistent with the true (but unknown) denoising steps.

The Final ELBO Expression

Combining all the terms, the final, more interpretable form of the ELBO (as shown at 04:03) is:

$$L_{DDPM} = \mathbb{E}_{q(x_1|x_0)}[\log p_\theta(x_0|x_1)] - D_{KL}(q(x_T|x_0)||p(x_T)) - \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)} \left[D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))\right]$$

This can be broken down into three distinct objectives: * **Reconstruction Term (L_0):** Maximize the log-likelihood of reconstructing the original data from the first latent state. * **Prior Matching Term (L_T):** Ensure the final latent state matches a simple prior distribution. * **Consistency Term ($L_{1:T-1}$):** Ensure the learned reverse (denoising) steps match the true reverse steps for all intermediate states.

Estimating the KL Divergence in the Consistency Term

The main challenge in optimizing the ELBO is computing the consistency term, specifically the KL divergence $D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))$. To do this, we first need an analytical expression for the posterior distribution $q(x_{t-1}|x_t, x_0)$.

Computing the Posterior $q(x_{t-1}|x_t, x_0)$

At (25:40), the instructor begins the derivation of this posterior.

1. Applying Bayes' Rule: We can express the desired posterior using Bayes' rule:

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

Due to the Markov property of the forward process, $q(x_t|x_{t-1}, x_0) = q(x_t|x_{t-1})$. This gives:

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

The term $q(x_t|x_0)$ is dropped as it's a constant with respect to x_{t-1}.

2. Using the Forward Process Definitions: We know the analytical forms for the distributions on the right-hand side because they are part of our defined forward process. * $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I)$ * $q(x_{t-1}|x_0) = \mathcal{N}(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}}x_0, (1-\bar{\alpha}_{t-1})I)$

Since the product of Gaussian PDFs is proportional to another Gaussian PDF, we know $q(x_{t-1}|x_t, x_0)$ is Gaussian. Its PDF is proportional to:

$$\exp\left(-\frac{1}{2}\left[\frac{(x_t - \sqrt{\alpha_t}x_{t-1})^2}{1-\alpha_t} + \frac{(x_{t-1} - \sqrt{\bar{\alpha}_{t-1}}x_0)^2}{1-\bar{\alpha}_{t-1}}\right]\right)$$

3. Completing the Square: The next step is to rearrange the terms in the exponent to match the standard quadratic form of a Gaussian in x_{t-1}, which is $-\frac{1}{2\sigma^2}(x_{t-1} - \mu)^2$. By expanding the squares and collecting terms for x_{t-1}^2 and x_{t-1}, we can identify the mean and variance of the posterior.

After a detailed algebraic manipulation (skipped in the video but essential for a full understanding), we arrive at the parameters of the posterior distribution $q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I)$.

The mean $\tilde{\mu}_t$ and variance $\tilde{\beta}_t$ are (as shown at 32:04):

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

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

Simplifying the KL Divergence

Now we have the two distributions needed for the KL divergence: 1. **True Posterior:** $q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I)$ 2. **Learned Model:** $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$

The KL divergence between two Gaussians $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$ has a closed form. A key design choice in DDPMs, mentioned at (35:11), is to **fix the variance of the learned model** to be equal to the true posterior's variance:

$$\Sigma_\theta(x_t, t) = \tilde{\beta}_t I$$

This is a reasonable choice because $\tilde{\beta}_t$ does not depend on the data x_0 and can be pre-computed. With this simplification, the KL divergence term reduces to a scaled mean-squared error between the means:

$$D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) = \frac{1}{2\tilde{\beta}_t}||\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)||_2^2 + C$$

where C is a constant that does not depend on θ.

This is the central result. The objective of matching the reverse process distributions simplifies to making the neural network $\mu_\theta(x_t, t)$ predict the true posterior mean $\tilde{\mu}_t(x_t, x_0)$.

Key Mathematical Concepts

Final DDPM Objective Function

The lecture culminates in presenting the final, simplified objective function for training a DDPM. The full ELBO is:

$$L = L_0 - L_T - \sum_{t=2}^{T} L_{t-1}$$

Where: * $L_0 = \mathbb{E}_q[\log p_\theta(x_0|x_1)]$ (Reconstruction Term) * $L_T = D_{KL}(q(x_T|x_0)||p(x_T))$ (Prior Matching Term - often ignored) * $L_{t-1} = \mathbb{E}_{q(x_t,x_0)}\left[\frac{1}{2\tilde{\beta}_t}||\tilde{\mu}_t(x_t,x_0) - \mu_\theta(x_t,t)||_2^2\right]$ (Consistency Term)

The training objective is to **maximize** this ELBO, which is equivalent to **minimizing** the negative ELBO. The most significant part of this objective is the consistency term, which trains the neural network to predict the mean of the true denoising step.

Recursive Property of the Forward Process

A crucial property, explained around (22:13), allows for efficient training. By unrolling the recursion of the forward process:

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1}$$

We can derive a direct relationship between x_t and x_0:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and $\epsilon \sim \mathcal{N}(0, I)$.

This implies that the distribution $q(x_t|x_0)$ is a Gaussian:

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

This is a powerful result because it means we can sample a noised image x_t for any timestep t in a single step, without having to iterate through the entire chain. This makes training highly efficient.

Visual Elements from the Video

The lecture is primarily based on handwritten derivations on a digital whiteboard. Key visual elements include:

- **ELBO Decomposition (00:42 - 02:04):** The instructor writes down the full ELBO and then rewrites the third term by applying the law of iterated expectations, visually separating the inner and outer expectations.
- **Consistency Term Breakdown (33:38):** A diagrammatic explanation of the consistency term, showing the "known denoising" distribution $q(x_{t-1}|x_t, x_0)$ and the "learnable denoising" distribution $p_\theta(x_{t-1}|x_t)$. This highlights that the goal is to match the learnable process to the known one.
- **Recursive Forward Process (22:13):** The instructor shows the recursive formula for x_t and then writes the final closed-form solution, emphasizing its importance for efficient sampling.

Self-Assessment for This Video

1. **Conceptual Question:** Explain in your own words the three main components of the simplified DDPM ELBO. Why is the "consistency term" the most important for training?

2. **Derivation:** Starting from Bayes' rule, derive the expression for the posterior distribution $q(x_{t-1}|x_t, x_0)$. Show that it is a Gaussian and identify its mean and variance.
3. **Simplification:** Why can the KL divergence term $D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))$ be simplified to a mean-squared error between the means of the two distributions? What assumption is made to achieve this?
4. **Efficiency:** What is the "nice property" of the forward process that allows for efficient training? Write down the formula for sampling x_t directly from x_0 and explain what each term represents.
5. **Comparison:** How do the terms in the DDPM ELBO relate to the terms in the standard VAE ELBO? What are the key similarities and differences?

Key Takeaways from This Video

- The ELBO for DDPMs can be elegantly decomposed into a reconstruction term, a prior matching term, and a sum of consistency terms over all timesteps.
- The training objective simplifies to making the learned reverse (denoising) process p_θ match the true (but analytically derived) reverse process q.
- By parameterizing the reverse process as a Gaussian and fixing its variance, the KL divergence objective becomes a simple and stable mean-squared error loss on the predicted mean.
- The forward process has a closed-form solution, allowing for efficient sampling of x_t at any timestep t directly from x_0, which is crucial for practical training.
- The entire framework, while mathematically intensive, results in a surprisingly simple and elegant final objective function that is effective for training deep generative models.