Study Material - Youtube

Document Information

- **Generated:** 2025-08-26 07:09:11
- **Source:** https://www.youtube.com/watch?v=2Sp0BqAWWXY
- **Platform:** Youtube
- **Word Count:** 1,651 words
- **Estimated Reading Time:** ~8 minutes
- **Number of Chapters:** 3
- **Transcript Available:** Yes (analyzed from video content)

Table of Contents

Video Overview

This lecture, "DDPMs as score-predictors," provides a profound alternative interpretation of Denoising Diffusion Probabilistic Models (DDPMs). The instructor demonstrates that beyond being simple noise predictors, DDPMs can be understood as models that learn the **score function** of the data distribution. This connection is established through a classical statistical result known as **Tweedie's formula**. By reframing the DDPM objective, the lecture shows that training a model to predict the added noise is mathematically equivalent to training it to predict the score. This insight is not only theoretically elegant, unifying DDPMs with score-based generative models, but also has significant practical implications, especially for conditional image generation.

Learning Objectives

Upon completing this lecture, students will be able to: - **Define and understand the Score Function** in the context of probability distributions. - **Explain Tweedie's Formula** and its relationship between a distribution's mean, variance, and score function. - **Connect DDPMs to Score-Based Models** by interpreting the DDPM as a score-predictor. - **Derive the relationship** between the true score of the noised data distribution and the noise added during the forward process. - **Recognize the mathematical equivalence** of different DDPM training objectives, such as noise prediction and score prediction.

Prerequisites

To fully grasp the concepts in this lecture, students should have a solid understanding of: - **Denoising Diffusion Probabilistic Models (DDPMs):** Familiarity with the forward (diffusion) and reverse (denoising) processes. - **Probability and Statistics:** Concepts of Gaussian distributions, conditional expectation, log-likelihood, and probability density functions. - **Calculus:** A firm grasp of vector calculus, particularly the gradient operator (∇). - **Previous DDPM Interpretations:** Knowledge that DDPMs can be trained to predict the original data x_0 or the added noise ϵ_t.

Key Concepts Covered

- Denoising Diffusion Probabilistic Models (DDPMs)
- Score Function
- Tweedie's Formula
- Score-Based Generative Modeling

- Equivalence of Training Objectives

DDPMs as Score-Predictors: A Deeper Interpretation

The lecture introduces a powerful and useful interpretation of DDPMs: viewing them as **score-predictors**. This perspective builds a bridge between DDPMs and another significant class of generative models known as score-based models.

The Score Function and Tweedie's Formula

Intuitive Foundation: What is a Score Function?

Before diving into the mathematics, let's build an intuition for the "score function."

> **Intuition:** Imagine a landscape where the height at any point represents the probability (or more accurately, the log-probability) of a data sample existing at that location. The **score function** at any point is simply the gradient of this landscape. It's a vector that points in the direction of the steepest ascent, i.e., towards regions where data is more probable.

In machine learning, we typically compute gradients with respect to model parameters (θ) to update our model. The score function is different; it is the gradient of the log-probability density function with respect to the **data variable** itself.

The score function for a probability distribution $p(t)$ is formally defined as:

$$\text{Score}(t) = \nabla_t \log p(t)$$

This function is fundamental to score-based generative modeling and, as we will see, is implicitly learned by DDPMs.

Mathematical Analysis: Tweedie's Formula

At the heart of the connection between DDPMs and score-based models is a classical result from statistics known as **Tweedie's Formula** (01:21). This formula provides a remarkable link between the mean, variance, and score function of a Gaussian distribution.

Theorem (Tweedie's Formula): Suppose a random variable t is drawn from a Gaussian distribution with mean μ_t and covariance Σ_t. The probability density function is $p(t) = \mathcal{N}(t; \mu_t, \Sigma_t)$. Tweedie's formula states that the conditional expectation of the mean μ_t, given an observation t, is:

$$\mathbb{E}[\mu_t | t] = t + \Sigma_t \cdot \nabla_t \log p(t)$$

Explanation of Terms: - $\mathbb{E}[\mu_t|t]$: The expected value (our best guess) of the distribution's mean, given that we have observed a sample t. - t: The observed data point. - Σ_t: The covariance matrix of the distribution. It acts as a scaling factor. - $\nabla_t \log p(t)$: The **score function** of the distribution $p(t)$.

This formula tells us that we can estimate the mean of a Gaussian distribution by taking an observation t and adjusting it by moving along the direction of the score function, scaled by the variance.

```
graph TD
    A["Observation<br/><i>t</i>"] --> C{"Tweedie's Formula<br/><b>E[ <sub>t</sub>|t] = t + Σ<sub>t</sub>
    B["Score Function<br/><i> <sub>t</sub>log p(t)</i><br/>(Direction of higher probability)"] --> C
    D["Covariance<br/><i>Σ<sub>t</sub></i><br/>(Scaling factor)"] --> C
    C --> E["Estimated Mean<br/><i>E[ <sub>t</sub>|t]</i>"]
```

Figure 1: A conceptual map illustrating the components of Tweedie's Formula.

Connecting DDPMs and Score Prediction

We can now apply Tweedie's formula to the DDPM framework to reveal its connection to score prediction.

Recalling the DDPM Forward Process

In the DDPM forward process, the distribution of the noisy data x_t at timestep t, given the original data x_0, is a Gaussian (05:03):

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

Here, we can map the DDPM terms to the variables in Tweedie's formula: - The observed data t is our noisy sample x_t. - The mean μ_t is $\sqrt{\bar{\alpha}_t}x_0$. - The covariance Σ_t is $(1-\bar{\alpha}_t)I$. - The distribution $p(t)$ is the marginal distribution of the noisy data, $p(x_t)$.

Deriving the True Score Function

By applying Tweedie's formula and rearranging terms, we can find an expression for the score of the marginal data distribution $p(x_t)$.

1. **Apply Tweedie's Formula (06:05):** The best estimate for the conditional mean $\mathbb{E}[\sqrt{\bar{\alpha}_t}x_0|x_t]$ is the true mean itself, $\sqrt{\bar{\alpha}_t}x_0$. Plugging this and the DDPM parameters into Tweedie's formula gives:

$$\sqrt{\bar{\alpha}_t}x_0 = x_t + (1-\bar{\alpha}_t)\nabla_{x_t}\log p(x_t)$$

2. **Relate x_0 to x_t and Noise ϵ_t (16:17):** From the definition of the forward process, $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t$. We can rearrange this to express x_0:

$$x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t)$$

3. **Substitute and Solve for the Score (17:25):** Substitute the expression for x_0 into the equation from step 1:

$$\sqrt{\bar{\alpha}_t}\left(\frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t)\right) = x_t + (1-\bar{\alpha}_t)\nabla_{x_t}\log p(x_t)$$
$$x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t = x_t + (1-\bar{\alpha}_t)\nabla_{x_t}\log p(x_t)$$
$$-\sqrt{1-\bar{\alpha}_t}\epsilon_t = (1-\bar{\alpha}_t)\nabla_{x_t}\log p(x_t)$$

Finally, solving for the score function, we get a crucial result:

$$\nabla_{x_t}\log p(x_t) = -\frac{\epsilon_t}{\sqrt{1-\bar{\alpha}_t}}$$

Key Insight: This elegant result (17:45) shows that the **true score of the noisy data distribution is simply the negatively scaled noise vector** ϵ_t that was used to generate x_t.

The DDPM Objective as Score Matching

This insight allows us to reinterpret the entire DDPM training process.

- **Noise Prediction Objective:** We previously established that DDPMs are trained to minimize the L2 distance between the true noise and the predicted noise:

$$L_{noise} \propto ||\epsilon_t - \hat{\epsilon}_\theta(x_t, t)||_2^2$$

- **Score Prediction Objective (14:03):** Using the relationship we just derived, we can see that this is equivalent to minimizing the L2 distance between the true score and a predicted score:

  $$L_{score} \propto ||\nabla_{x_t} \log p(x_t) - S_\theta(x_t)||_2^2$$

  where the predicted score $S_\theta(x_t)$ is related to the predicted noise $\hat{\epsilon}_\theta(x_t, t)$ by the same scaling factor:

  $$S_\theta(x_t) = -\frac{\hat{\epsilon}_\theta(x_t, t)}{\sqrt{1 - \bar{\alpha}_t}}$$

This means that a DDPM trained to predict noise is implicitly learning to predict the score function.

```
flowchart TD
    subgraph "DDPM as a Score Predictor"
        A["Input<br/>Noisy Data x<sub>t</sub>, Timestep t"] --> B{U-Net<br><i>S<sub> </sub>(x<sub>t</sub
        B --> C["Output<br/>Predicted Score S<sub> </sub>(x<sub>t</sub>)"];
        D["True Score<br/> <sub>x<sub>t</sub></sub>log p(x<sub>t</sub>) = - <sub>t</sub> / √(1-̄<sub>t</s
        C --> E;
    end
```

Figure 2: A flowchart illustrating the DDPM training process interpreted as regression over the score function. The model learns to predict the score of the data distribution at different noise levels.

Key Takeaways from This Video

- **DDPMs are Implicit Score-Predictors (21:50):** The central message is that training a DDPM to denoise an image (by predicting the added noise ϵ_t) is mathematically equivalent to training it to predict the score function ($\nabla_{x_t} \log p(x_t)$) of the noisy data distribution.
- **The True Score is Scaled Noise (18:07):** For the Gaussian noise schedule used in DDPMs, the true score is simply the negatively scaled version of the noise vector that was added. This provides a ground truth for the score that the model can learn.
- **Unification of Generative Models:** This interpretation unifies DDPMs with the family of score-based generative models, showing they are two sides of the same coin.
- **Practical Equivalence:** All common DDPM training objectives—predicting the original image x_0, predicting the added noise ϵ_t, or predicting the score—are equivalent up to scaling and shifting. Predicting the noise is often the most stable and widely used approach in practice.

Self-Assessment for This Video

Test your understanding of the concepts covered in this lecture with the following questions.

1. **Conceptual Understanding:**
   - In your own words, what is a “score function” and what does it represent intuitively?
   - Explain the core idea behind Tweedie’s formula. What three quantities does it connect?
   - What is the most important implication of the fact that the true score is proportional to the added noise in a DDPM?
2. **Mathematical Derivation:**
   - Given the DDPM forward process $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)I)$, write down the expression for the true score, $\nabla_{x_t} \log p(x_t)$, in terms of the added noise ϵ_t.
   - If a neural network $\hat{\epsilon}_\theta(x_t, t)$ predicts the noise, how would you define the corresponding predicted score function $S_\theta(x_t)$?
3. **Application and Interpretation:**

- Why are all the different interpretations of the DDPM objective (predicting noise, data, or score) considered mathematically equivalent?
- The instructor mentions this score-based view is particularly useful for conditional generation. Why might knowing the gradient of the log-probability be helpful in guiding the generation process?