Study Material - Youtube

Document Information

- **Generated:** 2025-08-26 07:04:08
- **Source:** https://www.youtube.com/watch?v=rNxtbFa8J-s
- **Platform:** Youtube
- **Word Count:** 2,260 words
- **Estimated Reading Time:** ~11 minutes
- **Number of Chapters:** 5
- **Transcript Available:** Yes (analyzed from video content)

Table of Contents

Video Overview

This video tutorial serves as a supplement to the Week 8 lectures on the "Mathematical Foundations of Generative AI." The primary goal is to provide detailed mathematical proofs for three fundamental concepts related to Normal (Gaussian) distributions, which are crucial for understanding Denoising Diffusion Probabilistic Models (DDPMs). These concepts were previously stated as facts, and this tutorial provides the rigorous derivations.

Learning Objectives

Upon completing this study material, you will be able to: - **Prove the distribution of the sum of two independent Normal random variables** using characteristic functions. - **Understand the role and properties of characteristic functions** in probability theory. - **Derive the formula for the Kullback-Leibler (KL) divergence** between two univariate Normal distributions from first principles. - **Recognize the key algebraic manipulations** required for working with Gaussian distributions, such as completing the square and handling expectations of quadratic forms. - **Appreciate the mathematical rigor** that underpins the algorithms used in modern generative models.

Prerequisites

To fully grasp the concepts in this video, you should have a solid understanding of: - **Probability Theory**: Random variables, probability density functions (PDFs), expectation, variance, and independence. - **Calculus**: Univariate integration and properties of the exponential function. - **Normal (Gaussian) Distribution**: Familiarity with its PDF, mean, and variance. - **Logarithm Properties**: Basic rules of logarithms, such as $\log(a/b) = \log(a) - \log(b)$.

Key Concepts Covered

1. **Sum of Two Independent Normal Random Variables**: Proving that the sum of two independent Gaussian variables is also a Gaussian variable.
2. **Characteristic Function of a Normal Distribution**: Using this powerful tool to prove properties of distributions.
3. **Kullback-Leibler (KL) Divergence**: A measure of difference between two probability distributions.

4. **Derivation of KL Divergence for Normal Distributions**: A step-by-step algebraic derivation of the KL divergence formula.

---

1. Sum of Two Independent Normal Random Variables

The first proof addresses a fundamental property of Gaussian distributions: their stability under addition. This means that when you add two independent Gaussian random variables, the result is another Gaussian random variable.

Intuitive Foundation

Imagine two separate processes, each with some uncertainty that can be modeled by a bell curve (a Gaussian distribution). For example, one machine cuts rods to a certain length with some random error, and another machine adds a cap of a certain length, also with some random error. If we want to know the distribution of the total length of the capped rod, we are essentially asking for the distribution of the sum of two random variables.

Intuitively: - The **new average length** (mean) should be the sum of the average rod length and the average cap length. - The **new total uncertainty** (variance) should be the sum of the individual uncertainties, as the errors from the two machines are independent.

This is precisely what the theorem states. The proof provides the mathematical justification for this intuition.

```
graph TD
    subgraph "Inputs"
        A["Random Variable X<br/>X ~ N( _x,  _x²)<br/>(e.g., Rod Length)"]
        B["Random Variable Y<br/>Y ~ N( _y,  _y²)<br/>(e.g., Cap Length)"]
    end

    subgraph "Process"
        C["Summation<br/>Z = X + Y<br/>(X and Y are independent)"]
    end

    subgraph "Output"
        D["Random Variable Z<br/>Z ~ N( _x +  _y,  _x² +  _y²)<br/>(e.g., Total Length)"]
    end

    A --> C
    B --> C
    C --> D
```

Figure 1: Conceptual flow of adding two independent Normal random variables.

Mathematical Analysis: Proof using Characteristic Functions

The instructor uses a powerful method involving **characteristic functions** to prove this property.

What is a Characteristic Function?

A characteristic function uniquely defines a probability distribution. It is the expectation of e^{itX} for a random variable X, where t is a real number and i is the imaginary unit.

$$\phi_X(t) = E[e^{itX}]$$

Two Key Properties: 1. **Uniqueness**: If two random variables have the same characteristic function, they have the same probability distribution. 2. **Sum of Independent Variables**: The characteristic function of a sum of independent random variables is the product of their individual characteristic functions. For independent X and Y, $\phi_{X+Y}(t) = \phi_X(t) \cdot \phi_Y(t)$.

Our strategy is to find the characteristic function of $Z = X + Y$ and show that it matches the known characteristic function of a Normal distribution.

Step-by-Step Derivation

(06:31) Step 1: State the Characteristic Function of a Normal Distribution

For a random variable $W \sim \mathcal{N}(\mu, \sigma^2)$, its characteristic function is a known result:

$$\phi_W(t) = \exp\left(it\mu - \frac{\sigma^2 t^2}{2}\right)$$

* The term $it\mu$ relates to the mean (a shift in the domain corresponds to a phase shift in the frequency domain). * The term $-\frac{\sigma^2 t^2}{2}$ is a Gaussian function in the variable t, which reflects the fact that the Fourier transform of a Gaussian is another Gaussian. The variance σ^2 determines the width of this Gaussian.

(07:11) Step 2: Apply the Property of Sums of Independent Variables

We are given: - $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ - $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ - $Z = X + Y$, with X and Y being independent.

The characteristic function of Z is:

$$\phi_Z(t) = \phi_{X+Y}(t) = E[e^{it(X+Y)}] = E[e^{itX}e^{itY}]$$

Because X and Y are independent, the expectation of the product is the product of the expectations:

$$\phi_Z(t) = E[e^{itX}]E[e^{itY}] = \phi_X(t) \cdot \phi_Y(t)$$

(07:33) Step 3: Substitute and Combine

Using the formula from Step 1, we write the characteristic functions for X and Y:

$$\phi_X(t) = \exp\left(it\mu_x - \frac{\sigma_x^2 t^2}{2}\right)$$

$$\phi_Y(t) = \exp\left(it\mu_y - \frac{\sigma_y^2 t^2}{2}\right)$$

Now, we multiply them to get $\phi_Z(t)$:

$$\phi_Z(t) = \exp\left(it\mu_x - \frac{\sigma_x^2 t^2}{2}\right) \cdot \exp\left(it\mu_y - \frac{\sigma_y^2 t^2}{2}\right)$$

Using the property $e^a \cdot e^b = e^{a+b}$, we combine the exponents:

$$\phi_Z(t) = \exp\left(\left(it\mu_x - \frac{\sigma_x^2 t^2}{2}\right) + \left(it\mu_y - \frac{\sigma_y^2 t^2}{2}\right)\right)$$

(08:15) Step 4: Rearrange to Match the Standard Form

We group the terms by it and t^2:

$$\phi_Z(t) = \exp\left((it\mu_x + it\mu_y) - \left(\frac{\sigma_x^2 t^2}{2} + \frac{\sigma_y^2 t^2}{2}\right)\right)$$

Factoring out the common terms gives:

$$\phi_Z(t) = \exp\left(it(\mu_x + \mu_y) - \frac{(\sigma_x^2 + \sigma_y^2)t^2}{2}\right)$$

(08:43) Step 5: Conclusion

This final expression for $\phi_Z(t)$ perfectly matches the standard form of a characteristic function for a Normal distribution, where: - The new mean is $\mu_Z = \mu_x + \mu_y$. - The new variance is $\sigma_Z^2 = \sigma_x^2 + \sigma_y^2$.

By the uniqueness property, we have proven that:

$$Z = X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

Extension to Linear Combinations

(10:17) The instructor also notes a related property for a scaled random variable. If $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and a is a constant, then the distribution of aX is:

$$aX \sim \mathcal{N}(a\mu_x, a^2\sigma_x^2)$$

This follows from the properties of expectation and variance: $E[aX] = aE[X]$ and $\text{Var}(aX) = a^2\text{Var}(X)$. This property is essential in the recursive formulas of DDPMs.

---

2. KL Divergence Between Two Normal Distributions

The second proof derives the formula for the KL divergence between two univariate Gaussian distributions. This is a critical component in the loss function of many generative models, including Variational Autoencoders (VAEs) and DDPMs.

Intuitive Foundation

KL divergence, $D_{KL}(P||Q)$, measures the "distance" or discrepancy from a "true" distribution P to an approximating distribution Q. It quantifies the extra information (measured in nats or bits) needed to encode samples from P when using a code optimized for Q.

> **Important Note:** KL divergence is **not a true metric** because it is **not symmetric**, i.e., $D_{KL}(P||Q) \neq D_{KL}(Q||P)$. The order matters.

In the context of generative models, we often want to make a simple distribution (e.g., our model's output) match a more complex target distribution. The KL divergence serves as a loss term that is minimized when the two distributions become identical.

Mathematical Analysis

(11:35) We want to compute $D_{KL}(P||Q)$ for: - $P(x) = \mathcal{N}(x; \mu_1, \sigma_1^2)$ - $Q(x) = \mathcal{N}(x; \mu_2, \sigma_2^2)$

The definition of KL divergence is:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx = E_{x\sim P(x)}[\log p(x) - \log q(x)]$$

Step-by-Step Derivation

(14:20) Step 1: Write the Log-Probability Density Functions

The PDF for a general Normal distribution $\mathcal{N}(x; \mu, \sigma^2)$ is:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Taking the natural logarithm (ln, which the instructor denotes as `log`):

$$\log p(x) = \log\left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right) + \log\left(\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)\right) = -\frac{1}{2}\log(2\pi\sigma_1^2) - \frac{(x-\mu_1)^2}{2\sigma_1^2}$$

Similarly for $q(x)$:

$$\log q(x) = -\frac{1}{2}\log(2\pi\sigma_2^2) - \frac{(x-\mu_2)^2}{2\sigma_2^2}$$

(15:48) Step 2: Find the Difference of the Log-PDFs

$$\log p(x) - \log q(x) = \left(-\frac{1}{2}\log(2\pi\sigma_1^2) - \frac{(x-\mu_1)^2}{2\sigma_1^2}\right) - \left(-\frac{1}{2}\log(2\pi\sigma_2^2) - \frac{(x-\mu_2)^2}{2\sigma_2^2}\right)$$

Rearranging the terms:

$$= \frac{1}{2}\left(\log(2\pi\sigma_2^2) - \log(2\pi\sigma_1^2)\right) - \frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2}$$

Using $\log a - \log b = \log(a/b)$:

$$= \frac{1}{2}\log\left(\frac{2\pi\sigma_2^2}{2\pi\sigma_1^2}\right) + \frac{(x-\mu_2)^2}{2\sigma_2^2} - \frac{(x-\mu_1)^2}{2\sigma_1^2}$$

$$= \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{(x-\mu_2)^2}{2\sigma_2^2} - \frac{(x-\mu_1)^2}{2\sigma_1^2}$$

(19:33) Step 3: Take the Expectation with Respect to P

Now we compute $E_{x\sim P}[\log p(x) - \log q(x)]$. By linearity of expectation:

$$D_{KL}(P||Q) = E_P\left[\log\frac{\sigma_2}{\sigma_1}\right] + E_P\left[\frac{(x-\mu_2)^2}{2\sigma_2^2}\right] - E_P\left[\frac{(x-\mu_1)^2}{2\sigma_1^2}\right]$$

Let's evaluate each term: 1. **First Term:** $\log(\sigma_2/\sigma_1)$ is a constant, so its expectation is itself.

$$E_P\left[\log\frac{\sigma_2}{\sigma_1}\right] = \log\frac{\sigma_2}{\sigma_1}$$

2. **Third Term:** This is the easiest. $E_P[(x-\mu_1)^2]$ is the definition of the variance of P, which is σ_1^2.

$$E_P\left[\frac{(x-\mu_1)^2}{2\sigma_1^2}\right] = \frac{1}{2\sigma_1^2}E_P[(x-\mu_1)^2] = \frac{1}{2\sigma_1^2}(\sigma_1^2) = \frac{1}{2}$$

3. **Second Term:** This requires more work.

$$E_P\left[\frac{(x-\mu_2)^2}{2\sigma_2^2}\right] = \frac{1}{2\sigma_2^2}E_P[(x-\mu_2)^2]$$

We need to compute $E_P[(x-\mu_2)^2]$. We know $E_P[x]=\mu_1$ and $\text{Var}_P(x)=\sigma_1^2$.

$$E_P[(x-\mu_2)^2] = E_P[((x-\mu_1)+(\mu_1-\mu_2))^2]$$
$$= E_P[(x-\mu_1)^2 + 2(x-\mu_1)(\mu_1-\mu_2) + (\mu_1-\mu_2)^2]$$
$$= E_P[(x-\mu_1)^2] + 2(\mu_1-\mu_2)E_P[x-\mu_1] + E_P[(\mu_1-\mu_2)^2]$$

- $E_P[(x-\mu_1)^2] = \text{Var}_P(x) = \sigma_1^2$. - $E_P[x-\mu_1] = E_P[x]-\mu_1 = \mu_1-\mu_1 = 0$. - $(\mu_1-\mu_2)^2$ is a constant, so its expectation is itself. Therefore:

$$E_P[(x-\mu_2)^2] = \sigma_1^2 + 0 + (\mu_1-\mu_2)^2$$

Substituting this back into the second term:

$$E_P\left[\frac{(x-\mu_2)^2}{2\sigma_2^2}\right] = \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2}$$

(25:18) Step 4: Combine All Terms

Putting everything together:

$$D_{KL}(P||Q) = \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

This is the final formula for the KL divergence between two univariate Normal distributions.

---

3. Posterior Distribution in DDPMs (Conceptual Outline)

(01:14) The instructor mentions a third key derivation from the DDPM lectures: finding the posterior distribution $q(x_{t-1}|x_t, x_0)$. He states this is an algebraic exercise using the "completing the square" technique and does not perform the derivation in this tutorial.

Context and Importance

In Denoising Diffusion Probabilistic Models (DDPMs), we have: - A **forward process** $q(x_t|x_{t-1})$ that gradually adds Gaussian noise to an image x_0 over T timesteps. - A **reverse process** $p_\theta(x_{t-1}|x_t)$ that learns to reverse this noising process, starting from pure noise x_T and generating an image x_0.

The training objective (the ELBO) requires us to compute the KL divergence between the true posterior $q(x_{t-1}|x_t, x_0)$ and our model's approximation $p_\theta(x_{t-1}|x_t)$. A remarkable property of the forward process is that this true posterior is tractable and can be shown to be a Gaussian distribution.

Derivation Outline (via Completing the Square)

1. **Apply Bayes' Theorem**:
$$q(x_{t-1}|x_t, x_0) \propto q(x_t|x_{t-1}, x_0) \cdot q(x_{t-1}|x_0)$$
(Note: $q(x_t|x_{t-1}, x_0) = q(x_t|x_{t-1})$ because the forward process is a Markov chain).
2. **Substitute Gaussian PDFs**: Each term on the right-hand side is a known Gaussian. We substitute their PDF expressions. The product of these Gaussians will have the form exp(quadratic in x_{t-1}).
3. **Complete the Square**: By collecting all terms involving x_{t-1}^2 and x_{t-1} in the exponent and algebraically manipulating them into the form $-\frac{(x_{t-1}-\tilde{\mu})^2}{2\tilde{\sigma}^2}$, we can identify the mean $\tilde{\mu}$ and variance $\tilde{\sigma}^2$ of the resulting Gaussian posterior.

This derivation, while algebraically intensive, is a cornerstone of DDPMs as it provides a computable target for the reverse process to learn.

Self-Assessment for This Video

1. **Question 1**: If $X \sim \mathcal{N}(10, 4)$ and $Y \sim \mathcal{N}(-2, 5)$ are independent, what is the distribution of $Z = X+Y$? What about $W = 3X$?
2. **Question 2**: Explain in your own words why the characteristic function is a useful tool for proving the sum of independent normal random variables is also normal.
3. **Question 3**: What is the key property of independent random variables that is used in the proof for the sum of normal variables? Write the corresponding mathematical expression.
4. **Question 4**: Calculate the KL divergence $D_{KL}(P||Q)$ where $P = \mathcal{N}(0, 1)$ and $Q = \mathcal{N}(1, 4)$.
5. **Question 5**: In the KL divergence formula $D_{KL}(P||Q)$, why is the expectation taken with respect to P and not Q? What does this imply about the asymmetry of the measure?

Key Takeaways from This Video

- **Stability of Normal Distributions**: The sum of independent Normal random variables is also a Normal random variable. The mean of the sum is the sum of the means, and the variance of the sum is the sum of the variances.
- **Power of Characteristic Functions**: They provide an elegant way to prove properties of distributions, especially for sums of independent variables, by transforming a convolution in the data domain into a simple multiplication in the frequency domain.
- **KL Divergence for Gaussians is Analytic**: The KL divergence between two Gaussian distributions has a closed-form analytical solution, which is crucial for its use as a loss function in machine learning.
- **Algebra is Key**: The derivations for these fundamental properties rely on careful and systematic algebraic manipulation, particularly with exponential and logarithmic functions and the "completing the square" technique.