Study Material - Youtube

Document Information

- **Generated:** 2025-08-26 06:17:56
- **Source:** https://www.youtube.com/watch?v=br7oydgTero
- **Platform:** Youtube
- **Word Count:** 1,833 words
- **Estimated Reading Time:** ~9 minutes
- **Number of Chapters:** 4
- **Transcript Available:** Yes (analyzed from video content)

Table of Contents

Video Overview

This video is a tutorial session from the IIT Madras B.S. Degree Programme, part of the "Mathematical Foundations of Generative AI" course. The lecture, delivered by Prof. Prathosh A P, focuses on providing a rigorous mathematical proof for **Jensen's Inequality**. The instructor highlights that this inequality is a cornerstone concept, particularly for the derivation of latent variable models, which are central to generative AI. The proof is demonstrated for the discrete case using the principle of mathematical induction.

Learning Objectives

Upon completing this study material, students will be able to: - **Understand and state** the formal definition of a convex function and Jensen's Inequality. - **Grasp the geometric intuition** behind convexity and what the inequality represents. - **Follow the step-by-step proof** of Jensen's Inequality for discrete random variables using mathematical induction. - **Appreciate the importance** of Jensen's Inequality as a foundational tool in machine learning and generative models.

Prerequisites

To fully understand the content of this lecture, students should be familiar with: - **Basic Calculus**: Concepts of functions, particularly functions from real numbers to real numbers ($\mathbb{R} \to \mathbb{R}$). - **Probability Theory**: Understanding of random variables (specifically discrete), probability distributions, and the concept of expectation ($\mathbb{E}[X]$). - **Summation Notation**: Comfort with using sigma ($\sum$) for summations. - **Proof Techniques**: Familiarity with proof by mathematical induction (base case, inductive hypothesis, and inductive step).

Key Concepts Covered

- Convex Functions
- Jensen's Inequality
- Proof by Mathematical Induction

Jensen's Inequality: A Foundational Concept

Intuitive Understanding of Convexity

(Timestamp: 02:16)

Before diving into the formal proof, it's crucial to understand the concept of a **convex function**.

Intuitive Idea: Imagine the graph of a function. If you pick any two points on the graph and draw a straight line (a chord) between them, a function is **convex** if this chord always lies *on or above* the function's graph. Visually, convex functions are "bowl-shaped". A classic example is the parabola $f(x) = x^2$.

Why it Matters for Jensen's Inequality: This "chord-above-graph" property is the geometric heart of Jensen's Inequality. The inequality compares the function's value at an average point with the average of the function's values. - **Function of an average:** f(average of x's) - **Average of a function:** average of $f(x)$'s

For a convex function, the function of an average is always less than or equal to the average of the function.

Mathematical Definition of a Convex Function

(Timestamp: 02:37)

The geometric intuition is formalized mathematically as follows.

A function $f : \mathbb{R} \to \mathbb{R}$ is defined as **convex** if for any two points x and y in its domain, and for any scalar $\lambda \in [0, 1]$, the following inequality holds:

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$$

Dissecting the Formula: - $\lambda x + (1-\lambda)y$: This is a **convex combination** of the points x and y. As λ varies from 0 to 1, this expression traces out all the points on the line segment between x and y. - $f(\lambda x + (1 - \lambda)y)$: This is the value of the function *on the curve* at a point between x and y. - $\lambda f(x) + (1 - \lambda)f(y)$: This is a convex combination of the function's values at x and y. This expression traces out all the points on the **chord** connecting the points $(x, f(x))$ and $(y, f(y))$ on the graph.

The inequality states that the point on the function's curve is always below or on the corresponding point on the chord.

```
graph TD
    subgraph Convex Function f(x)
        A((Point x))
        B((Point y))
        C["Convex Combination<br> x + (1- )y"]
        D["f( x + (1- )y)<br>Point on the curve"]
    end

    subgraph Chord between (x, f(x)) and (y, f(y))
        E["f(x)"]
        F["f(y)"]
        G["Convex Combination<br> f(x) + (1- )f(y)"]
        H["Point on the chord"]
    end

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

    A --> E
```

```
    B --> F
    E --> G
    F --> G
    G --> H

    D -- "is  " --> H
```

Caption: This diagram illustrates the components of the convexity definition. The value of the function at a convex combination of inputs (D) is less than or equal to the convex combination of the function's outputs (H).

Statement of Jensen's Inequality

(Timestamp: 01:05)

With the definition of convexity, we can now formally state Jensen's Inequality.

General Form: Let f be a convex function and X be a random variable. Jensen's Inequality states:

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

Intuition: The function of the average value of X is less than or equal to the average value of the function applied to X.

Discrete Form (for the proof): (Timestamp: 04:37) Let f be a convex function. For a set of points $\{x_1, x_2, \dots, x_n\} \in \mathbb{R}$ and a set of non-negative weights (or probabilities) $\{\alpha_1, \alpha_2, \dots, \alpha_n\}$ such that $\sum_{i=1}^{n} \alpha_i = 1$, the inequality is:

$$f\left(\sum_{i=1}^{n} \alpha_i x_i\right) \leq \sum_{i=1}^{n} \alpha_i f(x_i)$$

Here, $\sum \alpha_i x_i$ is the expectation of a discrete random variable that takes value x_i with probability α_i.

Proof of Jensen's Inequality (Discrete Case)

(Timestamp: 06:48)

The instructor proves the discrete version of Jensen's Inequality using **mathematical induction** on the number of points, n.

Proof Strategy: Mathematical Induction

The proof proceeds in three standard steps: 1. **Base Case:** Prove the statement is true for the smallest case, $n = 2$. 2. **Inductive Hypothesis:** Assume the statement is true for an arbitrary case, $n = k$. 3. **Inductive Step:** Using the assumption for $n = k$, prove that the statement is also true for $n = k + 1$.

```
flowchart TD
    A["<b>Start:</b> Prove for n=2<br/>(Base Case)"] --> B{"Is it true?"}
    B -->|Yes| C["<b>Assume:</b> True for n=k<br/>(Inductive Hypothesis)"]
    C --> D["<b>Prove:</b> True for n=k+1<br/>(Inductive Step)"]
    D --> E{"Is it true?"}
    E -->|Yes| F["<b>Conclusion:</b><br/>True for all n  2"]
    B -->|No| G["Proof Fails"]
    E -->|No| G
```

Caption: The logical flow of the proof by induction for Jensen's Inequality.

Step 1: Base Case (n=2)

(Timestamp: 07:07)

We need to show that the inequality holds for two points, x_1 and x_2, with weights α_1 and α_2 where $\alpha_1, \alpha_2 \geq 0$ and $\alpha_1 + \alpha_2 = 1$.

The inequality for $n = 2$ is:

$$f(\alpha_1 x_1 + \alpha_2 x_2) \leq \alpha_1 f(x_1) + \alpha_2 f(x_2)$$

Since $\alpha_1 + \alpha_2 = 1$, we can write $\alpha_2 = 1 - \alpha_1$. Let $\lambda = \alpha_1$, so $1 - \lambda = \alpha_2$. The inequality becomes:

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

This is precisely the **definition of a convex function**. Therefore, the base case is true by definition.

Step 2: Inductive Hypothesis (n=k)

(Timestamp: 07:55)

We **assume** that Jensen's inequality holds for any set of k points. For any $x_1, \dots, x_k$ and weights $\alpha_1, \dots, \alpha_k$ such that $\alpha_i \geq 0$ and $\sum_{i=1}^{k} \alpha_i = 1$, we assume:

$$f\left(\sum_{i=1}^{k} \alpha_i x_i\right) \leq \sum_{i=1}^{k} \alpha_i f(x_i)$$

Step 3: Inductive Step (n=k+1)

(Timestamp: 08:55)

Our goal is to prove the inequality for $k+1$ points. We start with the left-hand side (LHS) of the inequality for $n = k+1$:

$$\text{LHS} = f\left(\sum_{i=1}^{k+1} \alpha_i x_i\right)$$

where $\sum_{i=1}^{k+1} \alpha_i = 1$.

1. **Isolate the (k+1)-th term:**

$$\text{LHS} = f\left(\sum_{i=1}^{k} \alpha_i x_i + \alpha_{k+1} x_{k+1}\right)$$

2. **Rewrite the sum as a convex combination of two terms.** This is the key trick of the proof. We group the first k terms. Let $\beta = \alpha_{k+1}$. Then the sum of the first k weights is $\sum_{i=1}^{k} \alpha_i = 1 - \alpha_{k+1} = 1 - \beta$. We can rewrite the argument of f as:

$$\sum_{i=1}^{k+1} \alpha_i x_i = (1-\beta)\left(\sum_{i=1}^{k} \frac{\alpha_i}{1-\beta} x_i\right) + \beta x_{k+1}$$

3. **Apply the Base Case (n=2).** The expression is now a convex combination of two terms:

   - Term 1: $x_0 = \sum_{i=1}^{k} \frac{\alpha_i}{1-\beta} x_i$
   - Term 2: x_{k+1}
   - The weights are $(1-\beta)$ and β. Using the definition of convexity (our base case), we get:

$$f\left((1-\beta)x_0 + \beta x_{k+1}\right) \leq (1-\beta) f(x_0) + \beta f(x_{k+1})$$

4. **Apply the Inductive Hypothesis (n=k).** Now we look at the term $f(x_0)$. Let's define new weights $\tilde{\alpha}_i = \frac{\alpha_i}{1-\beta}$ for $i = 1, \dots, k$. These are valid weights because $\sum_{i=1}^{k} \tilde{\alpha}_i = \sum_{i=1}^{k} \frac{\alpha_i}{1-\beta} = \frac{1}{1-\beta} \sum_{i=1}^{k} \alpha_i = \frac{1-\beta}{1-\beta} = 1$. So, $x_0 = \sum_{i=1}^{k} \tilde{\alpha}_i x_i$. By our inductive hypothesis for k points:

$$f(x_0) = f\left(\sum_{i=1}^{k} \tilde{\alpha}_i x_i\right) \leq \sum_{i=1}^{k} \tilde{\alpha}_i f(x_i)$$

5. **Combine and Simplify.** Substitute the result from step 4 back into the inequality from step 3:

$$f\left(\sum_{i=1}^{k+1} \alpha_i x_i\right) \leq (1-\beta)\left(\sum_{i=1}^{k} \tilde{\alpha}_i f(x_i)\right) + \beta f(x_{k+1})$$

   Now, substitute back $\tilde{\alpha}_i = \frac{\alpha_i}{1-\beta}$:

$$\leq (1-\beta)\left(\sum_{i=1}^{k} \frac{\alpha_i}{1-\beta} f(x_i)\right) + \beta f(x_{k+1})$$

   The $(1-\beta)$ terms cancel out:

$$\leq \sum_{i=1}^{k} \alpha_i f(x_i) + \beta f(x_{k+1})$$

   Finally, substitute back $\beta = \alpha_{k+1}$:

$$\leq \sum_{i=1}^{k} \alpha_i f(x_i) + \alpha_{k+1} f(x_{k+1})$$

   This is exactly the right-hand side (RHS) for $n = k+1$:

$$\leq \sum_{i=1}^{k+1} \alpha_i f(x_i)$$

Since we have shown that if the statement is true for $n = k$, it is also true for $n = k+1$, and we have proven the base case for $n = 2$, by the principle of mathematical induction, **Jensen's Inequality is proven for all discrete random variables with a finite number of outcomes.**

Self-Assessment for This Video

1. **Conceptual Understanding:**
   - In your own words, what does it mean for a function to be convex? Draw a simple graph to illustrate.
   - Explain the core idea of Jensen's Inequality using the analogy of "function of an average" vs. "average of a function".
2. **Mathematical Formulation:**
   - Write down the mathematical definition of a convex function. What do the terms $\lambda x + (1-\lambda)y$ and $\lambda f(x) + (1-\lambda)f(y)$ represent geometrically?
   - State Jensen's inequality for a continuous random variable X with probability density function $p(x)$.
3. **Proof Analysis:**
   - Why is the base case ($n = 2$) in the proof of Jensen's Inequality considered "trivially true"?
   - What is the key algebraic manipulation performed in the inductive step (proving for $n = k+1$)? Explain why this step is necessary.

- In the proof, new weights $\tilde{\alpha}_i = \frac{\alpha_i}{1-\beta}$ are defined. Show mathematically why these new weights sum to 1.

4. **Application Problem:**
    - The function $f(x) = x^2$ is convex. Let a discrete random variable X take values $\{1, 5\}$ with equal probability $P(X = 1) = 0.5$ and $P(X = 5) = 0.5$. Verify that Jensen's Inequality, $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$, holds for this case.

Key Takeaways from This Video

- **Jensen's Inequality is a fundamental property of convex functions.** It establishes a relationship between the expectation of a function and the function of an expectation.
- **For a convex function f, $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$.** The inequality is reversed for concave functions.
- **The proof for the discrete case relies on mathematical induction.** The key is to cleverly apply the definition of convexity (the base case) and the inductive hypothesis to extend the proof from k to $k+1$ points.
- **This inequality is not just a mathematical curiosity; it is a critical tool** used in the derivation of many important results in machine learning, information theory, and optimization, including the Evidence Lower Bound (ELBO) for latent variable models like Variational Autoencoders (VAEs).