Study Material - Youtube

Document Information

- **Generated:** 2025-08-20 19:25:00
- **Source:** Based on IITM GenAI Curriculum W1L3
- **Platform:** Academic Content
- **Word Count:** 2,650 words
- **Estimated Reading Time:** ~12 minutes
- **Number of Chapters:** 3
- **Transcript Available:** No (compiled from academic sources)

Table of Contents

Video Overview

This lecture, "W1L3: f-Divergence," introduces the mathematical framework that underlies most modern generative models. f-Divergence provides a unified way to measure the "distance" between probability distributions, which is essential for training generative models. This lecture covers the general definition of f-divergence, explores important special cases like KL divergence and Jensen-Shannon divergence, and demonstrates their applications in generative adversarial networks (GANs) and variational autoencoders (VAEs).

Learning Objectives

Upon completing this lecture, a student will be able to: * **Define f-Divergence:** Understand the general mathematical formulation and properties of f-divergence measures. * **Identify Key Examples:** Recognize important special cases including KL divergence, reverse KL divergence, and Jensen-Shannon divergence. * **Apply to Generative Models:** Understand how different divergences are used in GANs, VAEs, and other generative frameworks. * **Compare Divergence Properties:** Analyze symmetry, boundedness, and computational properties of different divergences. * **Practical Implementation:** Compute divergences numerically and understand their optimization characteristics.

Prerequisites

To fully understand the concepts in this lecture, students should have: * **Probability Theory:** Solid understanding of probability distributions, probability density functions, and expectations. * **Information Theory:** Basic concepts of entropy and mutual information. * **Calculus:** Integration, differentiation, and optimization fundamentals. * **Convex Analysis:** Understanding of convex functions and Jensen's inequality. * **Machine Learning Basics:** Familiarity with loss functions and optimization objectives.

Key Concepts Covered

- f-Divergence Definition and Properties
- Kullback-Leibler (KL) Divergence
- Jensen-Shannon (JS) Divergence
- Total Variation Distance
- ^{2}-Divergence
- Applications in Generative Models

f-Divergence: The Mathematical Foundation for Generative Models

Introduction to Statistical Divergences

In generative modeling, we need to measure how different two probability distributions are. For example: - How close is our generated data distribution p_θ to the true data distribution p_{data}? - How can we quantify the "distance" between distributions mathematically? - What properties should a good distance measure have?

f-Divergence provides a unified mathematical framework for answering these questions.

General Definition of f-Divergence

Mathematical Formulation

For two probability distributions P and Q with densities $p(x)$ and $q(x)$, the **f-divergence** is defined as:

$$D_f(P\|Q) = \int q(x) f\left(\frac{p(x)}{q(x)}\right) dx$$

where $f : (0, \infty) \to \mathbb{R}$ is a **convex function** with $f(1) = 0$.

Alternative Discrete Formulation

For discrete distributions:

$$D_f(P\|Q) = \sum_i q_i f\left(\frac{p_i}{q_i}\right)$$

Key Properties

1. **Non-negativity:** $D_f(P\|Q) \geq 0$ for all P, Q
2. **Identity:** $D_f(P\|Q) = 0$ if and only if $P = Q$
3. **Convexity:** D_f is convex in both arguments
4. **Invariance:** Invariant under measure-preserving transformations

The Role of the Generator Function f

The choice of the convex function f determines the specific divergence: - **Domain:** $f : (0, \infty) \to \mathbb{R}$ - **Convexity:** Ensures non-negativity via Jensen's inequality - **Normalization:** $f(1) = 0$ ensures $D_f(P\|P) = 0$

Important Special Cases

1. Kullback-Leibler (KL) Divergence

Generator Function: $f(t) = t \log t$

Definition:

$$D_{KL}(P\|Q) = \int p(x) \log \frac{p(x)}{q(x)} dx = \mathbb{E}_{x\sim P}\left[\log \frac{p(x)}{q(x)}\right]$$

Discrete Version:

$$D_{KL}(P\|Q) = \sum_i p_i \log \frac{p_i}{q_i}$$

Key Properties: - **Asymmetric:** $D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$ in general - **Unbounded:** Can take values in $[0, \infty)$ - **Information-theoretic interpretation:** Expected log-likelihood ratio - **Used in:** Variational inference, VAEs, maximum likelihood estimation

Practical Interpretation: KL divergence measures the expected number of extra bits needed to encode samples from P when using a code optimized for Q.

2. Reverse KL Divergence

Generator Function: $f(t) = -\log t$

Definition:

$$D_{KL}(Q\|P) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

Key Difference from Forward KL: - **Forward KL** $D_{KL}(P\|Q)$: Mode-seeking, avoids placing mass where Q has little - **Reverse KL** $D_{KL}(Q\|P)$: Mode-covering, prefers to cover all modes of P

3. Jensen-Shannon (JS) Divergence

Generator Function: $f(t) = -(t+1)\log \frac{t+1}{2} + t \log t$

Definition:

$$D_{JS}(P\|Q) = \frac{1}{2} D_{KL}(P\|M) + \frac{1}{2} D_{KL}(Q\|M)$$

where $M = \frac{P+Q}{2}$ is the average distribution.

Key Properties: - **Symmetric:** $D_{JS}(P\|Q) = D_{JS}(Q\|P)$ - **Bounded:** $D_{JS}(P\|Q) \in [0, \log 2]$ (using natural logarithm) - **Square root metric:** $\sqrt{D_{JS}(P\|Q)}$ is a true metric - **Used in:** Original GANs, distribution comparison

4. Total Variation Distance

Generator Function: $f(t) = \frac{1}{2}|t-1|$

Definition:

$$D_{TV}(P\|Q) = \frac{1}{2} \int |p(x) - q(x)| dx = \frac{1}{2} \|P - Q\|_1$$

Key Properties: - **Symmetric:** $D_{TV}(P\|Q) = D_{TV}(Q\|P)$ - **Bounded:** $D_{TV}(P\|Q) \in [0, 1]$ - **Interpretation:** Maximum difference in probability assigned to any event - **Metric:** Satisfies triangle inequality

5. ^{2}-Divergence (Pearson ^{2})

Generator Function: $f(t) = (t-1)^2$

Definition:

$$D_{\chi^2}(P\|Q) = \int \frac{(p(x) - q(x))^2}{q(x)} dx$$

Key Properties: - **Asymmetric:** $D_{\chi^2}(P\|Q) \neq D_{\chi^2}(Q\|P)$ - **Unbounded:** Can take values in $[0, \infty)$ - **Statistical connection:** Related to ^{2}-test for goodness of fit

Variational Representation of f-Divergences

Fenchel-Legendre Conjugate

Every f-divergence has a **variational representation** based on the Fenchel-Legendre conjugate:

$$D_f(P\|Q) = \sup_T \left\{ \mathbb{E}_{x\sim P}[T(x)] - \mathbb{E}_{x\sim Q}[f^*(T(x))] \right\}$$

where f^* is the convex conjugate of f:

$$f^*(t) = \sup_u \{tu - f(u)\}$$

Example: KL Divergence Variational Form

For KL divergence ($f(t) = t \log t$, $f^*(t) = e^{t-1}$):

$$D_{KL}(P\|Q) = \sup_T \left\{ \mathbb{E}_{x\sim P}[T(x)] - \mathbb{E}_{x\sim Q}[e^{T(x)-1}] \right\}$$

This variational form is crucial for training GANs and other implicit generative models.

Applications in Generative Models

Generative Adversarial Networks (GANs)

Original GAN Objective: The original GAN paper showed that the optimal discriminator minimizes the Jensen-Shannon divergence:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]$$

Connection to JS Divergence: At the optimal discriminator, this becomes:

$$\min_G 2 \cdot D_{JS}(p_{\text{data}}\|p_g) - 2\log 2$$

f-GAN Framework

The f-GAN framework generalizes GANs to arbitrary f-divergences:

$$\min_G D_f(p_{\text{data}}\|p_g) = \min_G \sup_T \left\{ \mathbb{E}_{x\sim p_{\text{data}}}[T(x)] - \mathbb{E}_{x\sim p_g}[f^*(T(x))] \right\}$$

Different choices of f yield different GAN variants: - **KL-GAN:** Forward KL divergence - **Reverse-KL GAN:** Reverse KL divergence (mode-seeking) - **^{2}-GAN:** Least-squares GAN objective - **Total Variation GAN:** Wasserstein-like objective

Variational Autoencoders (VAEs)

VAEs optimize a variational lower bound that involves KL divergence:

$$\log p(x) \geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)\|p(z))$$

The KL term acts as a regularizer, encouraging the approximate posterior to stay close to the prior.

Computational Considerations

Numerical Implementation

KL Divergence (Discrete):

```
def kl_divergence(p, q, epsilon=1e-10):
    # Add small epsilon to avoid log(0)
    q_safe = np.maximum(q, epsilon)
    return np.sum(p * np.log(p / q_safe))
```

Jensen-Shannon Divergence:

```
def js_divergence(p, q):
    m = 0.5 * (p + q)
    return 0.5 * kl_divergence(p, m) + 0.5 * kl_divergence(q, m)
```

Optimization Properties

KL Divergence: - **Forward KL:** Mode-seeking, prefers sharp distributions - **Reverse KL:** Mode-covering, prefers broad distributions - **Asymmetry affects optimization behavior**

Jensen-Shannon Divergence: - **Symmetric:** More balanced optimization - **Bounded:** Stable gradients - **Can suffer from vanishing gradients when distributions are far apart**

Total Variation: - **Provides uniform upper bound on all probability differences** - **Can be difficult to optimize directly**

Relationships Between Divergences

Pinsker's Inequality

$$D_{TV}(P\|Q) \leq \sqrt{\frac{1}{2}D_{KL}(P\|Q)}$$

This connects KL divergence to total variation distance.

Variational Distance Relationships

For distributions that are not too far apart:

$$D_{TV}(P\|Q) \leq D_{JS}(P\|Q) \leq \frac{1}{2}D_{KL}(P\|Q)$$

f-Divergence Ordering

Different f-divergences can be ordered by their discriminatory power:

$$D_{TV} \leq D_{JS} \leq D_{KL}$$

(under appropriate conditions)

Choosing the Right Divergence

Factors to Consider

1. **Symmetry Requirements:**
   - Use JS or TV for symmetric comparison
   - Use KL when asymmetry is desired
2. **Boundedness:**
   - Bounded divergences (JS, TV) provide stable optimization
   - Unbounded divergences (KL) can provide stronger gradients
3. **Computational Efficiency:**
   - KL divergence is often easier to compute
   - Some divergences have closed-form solutions for specific distributions
4. **Mode-Seeking vs. Mode-Covering:**
   - Forward KL: mode-seeking
   - Reverse KL: mode-covering
   - JS: balanced behavior

Application-Specific Guidelines

For GANs: - JS divergence: Balanced training, can suffer from vanishing gradients - Wasserstein distance: More stable training, requires constraints

For VAEs: - KL divergence: Natural choice due to variational inference framework - -VAE: Weighted KL for disentanglement

For Model Selection: - Cross-entropy (related to KL): Standard choice for classification - JS divergence: Fair comparison of generative models

---

Key Takeaways from This Lecture

- **Unified Framework:** f-Divergence provides a mathematical foundation for measuring distances between probability distributions, unifying many important divergences under a single framework.
- **Generator Function:** The choice of the convex generator function f determines the specific divergence properties, including symmetry, boundedness, and optimization characteristics.
- **Key Examples:** Important special cases include KL divergence (asymmetric, unbounded, information-theoretic), Jensen-Shannon divergence (symmetric, bounded, used in GANs), and Total Variation distance (symmetric, bounded, metric).
- **Variational Representation:** All f-divergences can be expressed as variational optimization problems using Fenchel conjugates, enabling their use in adversarial training frameworks like GANs.
- **Generative Model Applications:** Different divergences lead to different generative model behaviors: KL divergence in VAEs for regularization, JS divergence in original GANs for balanced training, and f-GAN framework for exploring various divergences.
- **Optimization Properties:** The choice of divergence significantly affects training dynamics, with symmetric divergences providing more balanced optimization and bounded divergences ensuring stable gradients.
- **Practical Considerations:** Implementation requires careful handling of numerical stability, and the choice of divergence should be based on the specific requirements of symmetry, boundedness, and desired optimization behavior.

Self-Assessment for This Lecture

1. **Mathematical Definition:** Write the general definition of f-divergence and explain the role of the generator function f. What properties must f satisfy?
2. **Divergence Computation:** Calculate the KL divergence between two discrete distributions: $P = [0.7, 0.2, 0.1]$ and $Q = [0.3, 0.4, 0.3]$. Then compute $D_{KL}(Q\|P)$ and compare.
3. **Jensen-Shannon Derivation:** Show that the Jensen-Shannon divergence can be written as $D_{JS}(P\|Q) = H(M) - \frac{1}{2}[H(P) + H(Q)]$, where $M = \frac{P+Q}{2}$ and $H(\cdot)$ is entropy.
4. **GAN Connection:** Explain how the original GAN objective relates to Jensen-Shannon divergence. What are the advantages and disadvantages of this choice?
5. **Variational Representation:** Write the variational form of KL divergence using the Fenchel conjugate. How is this used in practice for training generative models?
6. **Comparison Analysis:** Compare KL divergence and Jensen-Shannon divergence in terms of: (a) symmetry, (b) boundedness, (c) optimization properties, (d) use cases in generative modeling.

7. **Mode-Seeking vs Mode-Covering:** Design a simple 2D example with a mixture of Gaussians to illustrate the difference between forward and reverse KL divergence optimization.

8. **f-GAN Framework:** For the generator function $f(t) = (t-1)^2$ (^{2}-divergence), derive the corresponding GAN objective. How does this relate to least-squares GAN?

9. **Numerical Implementation:** Implement a numerically stable version of JS divergence computation. What issues arise when probability values are very small?

10. **Application Selection:** For each scenario, choose the most appropriate divergence and justify your choice: (a) VAE training, (b) Fair comparison of two generative models, (c) Mode-covering generator training, (d) Stable GAN training with non-saturating gradients.