

Settings

Semantic search results
5

Prerequisite depth
2

Max concepts
10

Knowledge Graph Stats

Concepts in ChromaDB
3120

 Clear Chat

Erica - Your AI Tutor

Ask questions about AI/ML concepts from the Introduction to AI course



Can you explain the variational lower bound and how it relates to Jensen's inequality?



Understanding the Variational Lower Bound and Jensen's Inequality

1. Intuitive Overview

The variational lower bound, also known as the evidence lower bound (ELBO), is a key concept in variational inference, a method used to approximate complex probability distributions. In machine learning, particularly in probabilistic modeling and Bayesian inference, we often encounter intractable distributions that are difficult to compute directly. The variational lower bound provides a way to approximate these distributions by optimizing a simpler, tractable distribution.

Why It Matters:

- **Approximation of Intractable Distributions:** Many real-world problems involve complex distributions that are computationally infeasible to

work with directly. The variational lower bound allows us to approximate these distributions using simpler, more manageable distributions.

- **Optimization in Variational Inference:** The variational lower bound serves as an objective function that can be optimized to find the best approximation of the target distribution.
- **Connection to Jensen's Inequality:** Jensen's inequality is a fundamental mathematical result that underpins the derivation of the variational lower bound, providing a theoretical foundation for the approximation method.

2. Prerequisite Concepts

Before diving into the variational lower bound and Jensen's inequality, let's review some prerequisite concepts:

2.1. Probability Distributions

- **Probability Density Function (PDF):** A function that describes the relative likelihood for a continuous random variable to take on a given value.
- **Cumulative Distribution Function (CDF):** A function that gives the probability that a random variable is less than or equal to a certain value.

2.2. Expectation and Logarithms

- **Expectation ($E[\cdot]$):** The expected value of a random variable is the long-run average value of repetitions of the experiment it represents.
- **Logarithm (\log):** A mathematical function that is the inverse of exponentiation. It is used to simplify multiplicative processes into

additive ones.

2.3. Jensen's Inequality

- **Definition:** Jensen's inequality states that for a convex function f and a random variable X , the following holds:

$$f(E[X]) \leq E[f(X)]$$

- **Convex Function:** A function f is convex if for any x_1, x_2 in its domain and any $\lambda \in [0, 1]$,

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

- **Concave Function:** A function f is concave if $-f$ is convex.

3. Diving Deep into the Variational Lower Bound

3.1. Problem Setup

Consider a probabilistic model with latent variables Z and observed data X . The goal is to infer the posterior distribution $P(Z|X)$. However, this distribution is often intractable to compute directly.

3.2. Variational Inference

To approximate $P(Z|X)$, we introduce a variational distribution $Q(Z)$ that is easier to work with. The objective is to find $Q(Z)$ that is as close as possible to $P(Z|X)$.

3.3. KL Divergence

The Kullback-Leibler (KL) divergence is a measure of the difference between two probability distributions:

$$D_{\text{KL}}(Q(Z) \parallel P(Z|X)) = E_{Q(Z)} \left[\log \frac{Q(Z)}{P(Z|X)} \right]$$

Minimizing the KL divergence is equivalent to maximizing the variational lower bound.

3.4. Derivation of the Variational Lower Bound

The log marginal likelihood $\log P(X)$ can be rewritten using the variational distribution $Q(Z)$:

$$\log P(X) = \log \int P(X, Z) dZ$$

Using Jensen's inequality, we can derive the variational lower bound:

$$\log P(X) = \log \int P(X, Z) dZ \geq \int Q(Z) \log \frac{P(X, Z)}{Q(Z)} dZ$$

The right-hand side is the variational lower bound (ELBO):

$$\text{ELBO} = E_{Q(Z)} [\log P(X, Z)] - E_{Q(Z)} [\log Q(Z)]$$

3.5. Interpretation of the ELBO

- **First Term:** $E_{Q(Z)} [\log P(X, Z)]$ is the expected log-likelihood of the data under the joint distribution $P(X, Z)$.

- **Second Term:** $E_{Q(Z)} [\log Q(Z)]$ is the entropy of the variational distribution $Q(Z)$.

Maximizing the ELBO ensures that $Q(Z)$ is a good approximation of $P(Z|X)$.

4. Examples and Mathematical Details

4.1. Example: Gaussian Mixture Model (GMM)

Consider a Gaussian Mixture Model (GMM) with K components. The joint distribution $P(X, Z)$ is given by:

$$P(X, Z) = \prod_{i=1}^N P(x_i, z_i) = \prod_{i=1}^N \left(\sum_{k=1}^K \pi_k \mathcal{N}(x_i | \mu_k, \sigma_k^2) \right)$$

where π_k are the mixing coefficients, and $\mathcal{N}(x_i | \mu_k, \sigma_k^2)$ is the Gaussian distribution with mean μ_k and variance σ_k^2 .

The variational distribution $Q(Z)$ is typically chosen to be a factorized distribution:

$$Q(Z) = \prod_{i=1}^N Q(z_i)$$

where $Q(z_i)$ is a categorical distribution over the K components.

The ELBO can be written as:

$$\text{ELBO} = E_{Q(Z)} [\log P(X, Z)] - E_{Q(Z)} [\log Q(Z)]$$

$$= \sum_{i=1}^N E_{Q(z_i)} \left[\log \left(\sum_{k=1}^K \pi_k \mathcal{N}(x_i | \mu_k, \sigma_k^2) \right) \right] - \sum_{i=1}^N H(Q(z_i))$$

where $H(Q(z_i))$ is the entropy of $Q(z_i)$.

4.2. Example: Variance Gradient

The gradient of the log-likelihood with respect to the variance parameter σ_k^2 in a Gaussian component can be derived as:

$$\begin{aligned} \frac{\partial \log P(X, Z)}{\partial \sigma_k^2} &= \sum_{i=1}^N \frac{\partial}{\partial \sigma_k^2} \log \mathcal{N}(x_i | \mu_k, \sigma_k^2) \\ &= \sum_{i=1}^N \left(-\frac{1}{2\sigma_k^2} + \frac{(x_i - \mu_k)^2}{2(\sigma_k^2)^2} \right) \end{aligned}$$

This gradient is used in the optimization process to update the variance parameters.

5. Relevant Resources for Further Reading

- Textbooks and Papers:
 - [Dasgupta, Papadimitriou, Vazirani] Algorithm Design [Resource: <http://algorithmics.lsi.upc.edu/docs/Dasgupta-Papadimitriou-Vazirani.pdf>]
 - [LeCun, Bottou, Orr, Müller] Efficient BackProp [Resource: <http://yann.lecun.com/exdb/publis/pdf/lecun-98b.pdf>]
 - [Stanford CS224N] Lecture Notes on Word Vectors [Resource: <https://web.stanford.edu/class/cs224n/readings/cs224n-2019->

[notes01-wordvecs1.pdf](#)

- [Pascanu, Mikolov, Bengio] On the Difficulty of Training Recurrent Neural Networks [Resource: <http://proceedings.mlr.press/v28/pascanu13.pdf>]
- [Kingma, Welling] Auto-Encoding Variational Bayes [Resource: <https://arxiv.org/pdf/1503.04069v1.pdf>]
- Online Courses and Assignments:
 - [Pantelis] Variational Inference and Gaussian Mixture Models [Resource: <https://pantelis.github.io/aiml-common/assignments/topics/mle/gaussian/mle-gmm-sgd-preview.html>]
 - [Pantelis] Variance Gradient and Stochastic Gradient Descent [Resource: <https://pantelis.github.io/aiml-common/assignments/main/ai-fall-2025/assignment-1-grad.html>]
 - [Pantelis] AI Course Syllabus [Resource: <https://pantelis.github.io/courses/ai/syllabus/index.html>]
 - [Pantelis] Logistic Regression and CTR Prediction [Resource: <https://pantelis.github.io/aiml-common/assignments/topics/logistic-regression/ctr-preview.html>]
 - [Pantelis] Batch Normalization in Optimization [Resource: <https://pantelis.github.io/aiml-common/lectures/optimization/batch-normalization/index.html>]
- Video Tutorials:
 - [3Blue1Brown] The Essence of Calculus, Chapter 7: Derivatives of Exponentials [Resource: <https://www.youtube.com/watch?v=6niqTuYFZLQ&t=521s>]

6. Summary and Next Steps

Summary:

- **Variational Lower Bound (ELBO):** A method to approximate intractable posterior distributions in probabilistic models by optimizing a simpler, tractable distribution.
- **Jensen's Inequality:** A mathematical result that underpins the derivation of the variational lower bound, ensuring that the ELBO is a lower bound on the log marginal likelihood.
- **Examples:** The ELBO is used in variational inference for models like Gaussian Mixture Models, and its gradient is used to update parameters in optimization.

Next Steps:

- **Explore Variational Inference in More Depth:** Dive into more advanced topics such as mean-field approximation, reparameterization tricks, and amortized inference.
- **Study Specific Models:** Apply variational inference to more complex models like Variational Autoencoders (VAEs) and Deep Generative Models.
- **Optimization Techniques:** Learn about advanced optimization techniques for training variational models, such as Adam and RMSprop.

By understanding the variational lower bound and its connection to Jensen's inequality, you'll be well-equipped to tackle a wide range of probabilistic modeling and Bayesian inference problems. Keep exploring

and experimenting with different models and techniques to deepen your understanding!

▼  View Retrieved Context

Seed Concepts:

- Lower Bound (score: 0.509)
- Slack Variable (score: 0.462)
- Variance Gradient (score: 0.427)
- Inequality Constraints (score: 0.427)
- Min-Max Theorem (score: 0.426)

Explanation Order:

Ask a question about AI/ML... 

Resources: 34

Examples: 16