Mathematical Foundations of Bayesian Neural Networks

1 Introduction

Bayesian Neural Networks (BNNs) incorporate Bayesian probability principles, allowing for the estimation of uncertainty in predictions. This document outlines the mathematical framework underlying BNNs, focusing on priors, likelihood, and posterior distributions.

2 Mathematical Components

2.1 Priors

Priors in a BNN express initial beliefs about the values of the parameters θ (weights and biases) before any data is observed. A common choice is the Gaussian distribution:

$$p(\theta) = \mathcal{N}(\theta; \mu_0, \sigma_0^2)$$

where μ_0 and σ_0^2 are the mean and variance of the prior distribution, respectively.

2.2 Likelihood

The likelihood function $p(D|\theta)$ measures the probability of observing the data D given parameters θ. This varies depending on the task:

2.2.1 Regression Tasks

For regression tasks, the likelihood is typically modeled as:

$$p(D|\theta) = \prod_{i=1}^{N} \mathcal{N}(y_i; f(x_i, \theta), \sigma^2)$$

where y_i are the observed outputs, x_i are the inputs, f represents the neural network function parameterized by θ, and σ^2 is the noise variance.

2.2.2 Classification Tasks

For classification tasks, the likelihood is given by:

$$p(D|\theta) = \prod_{i=1}^{N} \text{Categorical}(y_i; \text{softmax}(f(x_i, \theta)))$$

where y_i are categorical labels.

2.3 Posterior

The posterior distribution $p(\theta|D)$ combines the prior and the likelihood through Bayes' Theorem:

$$p(\theta|D) = \frac{p(D|\theta) \times p(\theta)}{p(D)}$$

where $p(D)$ is the evidence, computed as:

$$p(D) = \int p(D|\theta) \times p(\theta)\, d\theta$$

This integral is generally intractable, leading to approximation methods such as Markov Chain Monte Carlo (MCMC) and Variational Inference (VI).

2.3.1 Markov Chain Monte Carlo (MCMC)

In MCMC, parameters are updated via:

$$\theta^{(t+1)} = \theta^{(t)} + \epsilon \cdot \nabla_\theta \log p(\theta^{(t)}|D)$$

where ϵ is a step size.

2.3.2 Variational Inference (VI)

In VI, the posterior is approximated by a simpler distribution $q(\theta)$:

$$q(\theta) \approx p(\theta|D)$$

Here, $q(\theta)$ is often chosen to be Gaussian, adjusted to minimize the KL divergence from the true posterior.