

---

# Improve Predictive Model Calibration for Safety-Critical Domains

---

**Le Van Tuan Anh\***

anh.1vt2416657@sis.hust.edu.vn

**Nguyen Tran Minh Tri\***

tri.ntm2400116@sis.hust.edu.vn

**Le Xuan Nhat Khoi\***

khoi.lxn2416711@sis.hust.edu.vn

**Nguyen Duy Tung\***

tung.nd2416762@sis.hust.edu.vn

## Abstract

Reliable uncertainty estimation is crucial for deploying modern neural networks in high-stakes scenarios, yet standard deterministic models trained with conventional loss functions often become systematically overconfident in their predictions. In this work, we propose a Bayesian neural network (BNN) framework that improves model calibration by replacing traditional point-estimate training with optimization of the evidence lower bound (ELBO). Our approach learns a distribution over network weights, enabling principled uncertainty quantification and mitigating the overconfidence commonly induced by cross-entropy-based training. Empirical results demonstrate that our BNN formulation yields significantly better calibration, measured via expected calibration error and negative log-likelihood, while maintaining competitive predictive accuracy.

## 1 Introduction

Neural networks have achieved remarkable performance across a wide range of predictive tasks, yet their reliability remains a significant challenge in practical deployment. Standard deterministic models trained with conventional loss functions, most notably cross-entropy, tend to produce overly confident predictions, even when faced with ambiguous inputs or samples far outside the training distribution. This systematic overconfidence not only impairs decision-making but also limits the applicability of modern deep learning systems in safety-critical domains such as medical diagnosis, autonomous navigation, and scientific modeling.

Bayesian neural networks (BNNs) offer a principled framework for addressing this limitation by replacing point estimates of network weights with full posterior distributions. Instead of committing to a single set of parameters, BNNs represent uncertainty directly through the variability of sampled weight configurations, enabling more calibrated predictive distributions. However, training BNNs remains challenging due to the intractability of the exact posterior, which necessitates efficient and scalable approximations.

In this work, we revisit variational Bayesian learning as a practical solution to these challenges. We propose a BNN framework trained via maximization of the evidence lower bound (ELBO), which balances model fit with posterior regularization through a Kullback–Leibler divergence term. Unlike traditional losses that focus solely on maximizing predictive accuracy, the ELBO encourages the model to capture epistemic uncertainty by constraining the posterior to remain close to a meaningful prior. This leads to better uncertainty estimates, improved robustness to distribution shift, and reduced overconfidence in regions where data are sparse or ambiguous.

---

\*DS-AI 03 K69

To assess the effectiveness of our approach, we conduct extensive empirical evaluations across classification tasks and calibration benchmarks. Our results show that ELBO-based training consistently improves calibration metrics, including expected calibration error and negative log-likelihood, while maintaining competitive accuracy. These findings highlight the value of Bayesian training objectives in producing trustworthy and informative predictions.

## 2 Bayesian Inference

Bayesian inference provides a principled mathematical framework for reasoning under uncertainty by treating model parameters as random variables rather than fixed quantities. Given observed data  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$  and model parameter  $\theta$ , Bayes' rule defines the posterior distribution as:

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

where  $p(\theta)$  is a prior encoding assumptions about the parameters,  $p(\mathcal{D} | \theta)$  is the likelihood of the data under the model, and  $p(\mathcal{D})$  is the marginal likelihood (the evidence). The posterior captures all uncertainty about the parameters after observing data, and predictions are obtained by marginalizing over this uncertainty:

$$p(y^* | x^*, \mathcal{D}) = \int p(y^* | x^*, \theta)p(\theta | \mathcal{D})d\theta$$

### 2.1 Bayesian Neural Networks

A Bayesian Neural Network (BNN) applies Bayesian inference to deep learning by placing distributions over the network weights instead of learning deterministic parameters. Let  $\theta$  represent all weights and biases of the network. In a BNN, we specify:

- A **prior**  $p(\theta)$ , typically chosen as a factorized Gaussian with zero mean
- A **likelihood**  $p(\mathcal{D} | \theta)$ , which depends on the model output (e.g., softmax likelihood for classification)

Because the neural network likelihood is highly nonlinear and the parameter space is high-dimensional, the posterior is **intractable**, we cannot compute it analytically or normalize it exactly. Likewise, exact predictive inference requires evaluating an integral over all possible weight values, which is computationally prohibitive.

### 2.2 Probabilistic Inference

To address this intractability, BNNs rely on approximate inference methods. The two primary families are:

- **Markov chain Monte Carlo (MCMC)**: Draws samples from the true posterior using iterative sampling schemes (e.g., Hamiltonian Monte Carlo). While theoretically accurate, MCMC scales poorly with deep architectures and large datasets.
- **Variational Inference (VI)**: Approximates the posterior by a simpler distribution  $q_\phi(\theta)$  and selects  $q_\phi(\theta)$  to minimize the divergence from the true posterior. VI is computationally efficient, scalable to large datasets, and compatible with gradient-based optimization.

Variational inference forms the foundation of most practical BNN implementations, as it enables amortized training through stochastic optimization and mini-batching.

### 2.3 Predictive Uncertainty

BNNs provide two forms of uncertainty:

- **Epistemic Uncertainty**: Arising from limited data or model ambiguity, captured through variability in  $p(\theta | \mathcal{D})$
- **Aleatoric Uncertainty**: Representing inherent noise in the data, captured through the likelihood  $p(y | x, \theta)$

**3 ELBO Loss as an Objective**

**4 Bayesian Linear and Bayesian Convolution**

**5 Experiments**

**6 Conclusion**