From Aleatoric to Epistemic: Exploring Uncertainty Quantification Techniques in Artificial Intelligence

Tianyang Wang, Yunze Wang, Jun Zhou et al.

April 11, 2025

Presented by Yuqi Li

Introduction

Key Problems in AI Uncertainty:

- Models often make overconfident predictions
- Traditional metrics don't assess uncertainty quality
- Need to distinguish between aleatoric (data) and epistemic (model) uncertainty

Why It Matters:

- Safety-critical applications (healthcare, autonomous vehicles)
- Better decision-making under uncertainty calibration and reliability

Types of Uncertainty

Aleatoric Uncertainty

- Inherent data noise
- Cannot be reduced with more data
- Example: Sensor noise in measurements

Epistemic Uncertainty

- Model uncertainty
- Can be reduced with more data
- Example: Limited training samples

Modern Techniques in Uncertainty Quantification

Bayesian Methods

- Bayesian Neural Networks
- Monte Carlo Dropout
- Variational Inference

Ensemble Methods

- Deep Ensembles
- Bootstrap Aggregating

Direct Uncertainty Prediction

- Evidential Deep Learning
- Conformal Prediction
- Quantile Regression

Bayesian Neural Networks (BNNs)

Core Idea:

- Place distributions over weights $\theta \sim p(\theta|D)$
- Predictions become probability distributions:

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

- Epistemic Uncertainty: variance in $p(\theta|D)$ reflects model uncertainty
- Aleatoric uncertainty: $p(y|x, D)$ captures noise

Challenges:

- Computationally expensive
- Approximate inference required

Variational Inference (VI)

The Core Idea:

- Approximate the true posterior $p(\theta \mid D)$ with a simpler distribution $q_\phi(\theta)$.
- Minimize KL divergence:

$$\mathrm{KL}\big(q_\phi(\theta) \parallel p(\theta \mid D)\big).$$

Key Benefits:

- Scalability: Works with deep networks.
- Reparameterization Trick:

$$\theta = \mu + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I).$$

- Theoretical Guarantees: Bounds the approximation error.

Evidence Lower Bound (ELBO)

$$\log p(D) \;\geq\; \mathbb{E}_q[\log p(D \mid \theta)] \;-\; \mathrm{KL}\big(q_\phi(\theta) \parallel p(\theta)\big).$$

Limitation: Mean-field assumption may underestimate uncertainty.

Monte Carlo Dropout (MC Dropout)

Practical Bayesian Approximation:

- Enable dropout at *test time*
- Forward passes with T different dropout masks
- T typically 20-100 samples

Why It Works:

- Implicit posterior sampling
- No additional parameters

Andvanced Techniques:

- Hamiltonian Monte Carlo (HMC): incorporates gradient information to explore the posterior more effectively
- Sequential Monte Carlo (SMC): updates posterior samples sequentially

Monte Carlo Dropout (MC Dropout)

Total Predictive Variance (Kendall and Gal (2017))

$$\underbrace{\mathbb{V}[y]}_{\text{Total}} = \underbrace{\mathbb{V}[\mathbb{E}[y|x, W]]}_{\text{Epistemic}} + \underbrace{\mathbb{E}[\mathbb{V}[y|x, W]]}_{\text{Aleatoric}}$$

- For classification (discrete outputs):

$$\text{Uncertainty} = \text{Var}\left(\frac{1}{T}\sum_{t=1}^{T} p(y|x, \theta_t)\right)$$

- For regression (continuous outputs):

$$\sigma^2 = \underbrace{\text{Var}(\mu_t)}_{\text{Epistemic}} + \underbrace{\frac{1}{T}\sum_{t=1}^{T}\sigma_t^2}_{\text{Aleatoric}}$$

where $(\mu_t, \sigma_t^2) = f(x; \theta_t)$

Generative Models for Uncertainty Quantification

Why Generative Models?

- Learn data distribution $p(x)$ directly
- Capture aleatoric uncertainty via density estimation
- Provide epistemic uncertainty through latent space analysis

Common Approaches:

- VAEs: Approximate posterior with ELBO

$$\mathcal{L} = \mathbb{E}_{q_\phi}[\log p_\theta(x|z)] - \mathrm{KL}(q_\phi \| p(z))$$

- GANs: Discriminator scores as uncertainty indicators
- Normalizing Flows: Exact density computation (next slides)

UQ Applications

- **Anomaly Detection**: Low $p(x)$ = high uncertainty
- **Prediction**: Confidence intervals via latent sampling
- **Active Learning**: Select low-density points

Generative Models for Uncertainty: Normalizing Flows

Idea:

- Use a series of **invertible** and **differentiable** transformations to map a simple distribution (e.g. $\mathcal{N}(0, I)$) to a complex target distribution.

Density Computation:

$$x = f(z) = f_K \circ f_{K-1} \circ \cdots \circ f_1(z).$$

$$p_X(x) = p_Z\left(f^{-1}(x)\right) \times \left|\det \frac{\partial f^{-1}(x)}{\partial x}\right|.$$

UQ Workflow:

- *Sampling and Inverse Mapping*: Draw z from base distribution, then $x = f(z)$.
- *Optimization*: Train flow to maximize $\log p(x)$ on normal data
- *Uncertainty*: High $p(x') \rightarrow$ Low uncertainty (in-distribution); Low $p(x') \rightarrow$ High uncertainty (OOD/anomaly)

Generative Models for Uncertainty: Normalizing Flows

Why Use Normalizing Flows?

- **Exact Likelihood**: Unlike GANs, can compute $\log p(x)$ and density-based uncertainty directly
- **Invertible**: Allows reconstruction and easy sampling, no variational approximation error
- **Broad Applications**: Density estimation, anomaly/OOD detection, variational inference, etc.

Challenges:

- Must maintain **invertibility** $\rightarrow$ architecture constraints.
- Potentially **high computational cost** for deep flows.

Popular Architectures:

- **RealNVP**: Coupling layers for efficient Jacobian computation.
- **Glow**: Extends RealNVP with *1x1* convolutions and ActNorm for high-res image modeling.
- **MAF / IAF**: Autoregressive designs for flexible density estimation and VAE enhancement.

Deep Ensembles

Non-Bayesian Alternative

- Train M models with different initializations, combine predictions:

$$p(y \mid x) = \frac{1}{M} \sum_{m=1}^{M} p(y \mid x, \theta_m)$$

- Captures both aleatoric and epistemic uncertainty

Advantages

- **Robustness**: Often resilient to adversarial perturbations
- **Good Empirical Performance**: Works well across various tasks (e.g. medical image diagnosis, autonomous navigation)
- **Simplicity**: Conceptually straightforward (no need for explicit Bayesian priors)

Deep Ensembles

Challenges

- High Computational Cost
- Memory Requirements
- **Underestimation of Uncertainty?** (if diversity among ensemble members is not truly large)

Potential Remedies

- **Distillation-based Ensemble Approximations**: Use knowledge distillation to compress the ensemble into a single model
- **Shared-weight Architectures**: Partially share parameters to reduce memory footprint

Bootstrap Aggregating (Bagging)

Data-Centric Ensembling:

- Train on bootstrap resamples of data, like data perturbation
- Variance indicates epistemic uncertainty, disagreement between models reflects knowledge gaps
- Bootstrap mimics Bayesian marginalization over datasets

$$\text{Uncertainty} = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(f_m(x) - \bar{f}(x))^2}$$

Key Advantages

- Works with any base learner (trees, NNs, etc.), particularly effective for Random Forests
- **No Distributional Assumptions**: Pure data-driven approach

Evidential Deep Learning (EDL)

Core Mechanism:

- Outputs evidence values $\mathbf{e}_k = f_\theta(x)_k \geq 0$ per class
- Forms Dirichlet distribution:

$$p(y|x) = \text{Dir}(\alpha), \quad \alpha = \mathbf{e} + 1$$

- Uncertainty decomposition:

$$\underbrace{u}_{\text{Epistemic}} = \frac{K}{\sum \alpha_k}, \quad \underbrace{\text{Var}(\alpha)}_{\text{Aleatoric}}$$

Key Features

- Single-pass inference: No MC sampling needed
- Regularization: Penalize small evidence to prevent overconfidence
- Applications: Medical diagnosis, autonomous driving

Distribution-Free Uncertainty Quantification

Conformal Prediction

- Construct prediction intervals via permutation
- Makes no distributional assumptions

For UQ:

1. Compute nonconformity scores $s(x_i, y_i)$
2. Find $\hat{q} = (1-\alpha)$-quantile of scores
3. Output interval:

$$C(x) = \{y : s(x, y) \leq \hat{q}\}$$

Guarantee:

$$P(y \in C(x)) \geq 1 - \alpha$$

Quantile Regression

- Directly model conditional quantiles
- Optimize with pinball loss

For UQ:

1. Train quantile models:

$$\{\hat{y}_\tau = f_\tau(x)\}_{\tau=\alpha/2}^{1-\alpha/2}$$

2. Build prediction interval:

$$[\hat{y}_{\alpha/2}, \hat{y}_{1-\alpha/2}]$$

Loss Function:

$$L_\tau = \max(\tau(y - \hat{y}_\tau), (1-\tau)(\hat{y}_\tau - y))$$

Key Advantages: Finite-sample validity, Model-agnostic

Comparison of Uncertainty Quantification Methods

Method	Aleatoric	Epistemic	Computational Cost	Best For
Bayesian Neural Nets (BNNs)	✓	✓	High	Small datasets, theoretical rigor
MC Dropout	✓	✓	Medium	Quick implementation, existing models
Deep Ensembles	✓	✓	High	State-of-the-art accuracy
Conformal Prediction	✓	(✓)	Low	Distribution-free guarantees
Evidential Deep Learning	✓	(✓)	Medium	Classification tasks

Key Insights

- Full Bayesian methods (BNNs) capture both uncertainties but are computationally expensive
- Approximate methods (MC Dropout, Evidential) offer better efficiency
- Conformal Prediction provides strongest theoretical guarantees

Note: (✓) indicates partial capability

Metrics: 1. Calibration

Calibration measures whether a model's confidence scores reflect true probabilities.
Expected Calibration Error (ECE):

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|$$

Maximum Calibration Error (MCE):

$$MCE = \max_{m \in \{1,\ldots,M\}} |\text{acc}(B_m) - \text{conf}(B_m)|$$

Brier Score (MSE):

$$BS = \frac{1}{n} \sum_{i=1}^{N} (p_i - y_i)^2$$

where p_i is predicted probability and y_i is actual outcome.

Metrics: 2. Sharpness Metrics

Sharpness assesses the concentration of the predictive distribution, independent of its calibration.

Prediction Interval Width (PIW): in regression tasks, PIW evaluates the sharpness of confidence intervals.

$$PIW = \frac{1}{n} \sum_{i=1}^{N} (U_i - L_i)$$

where U_i and L_i are upper/lower bounds of CI for ith sample.

Entropy: for classification tasks

$$H(p(y|x)) = -\sum_{k=1}^{K} p(y = k|x) \log p(y = k|x)$$

Metrics: 3. Scoring Rules

Scoring rules provide a unified framework to evaluate predictive distributions by combining calibration and sharpness into a single metric.

Logarithmic Score:

$$\text{Log Score} = -\frac{1}{n}\sum_{i=1}^{n} \log p(y_i|x_i)$$

Continuous Ranked Probability Score (CRPS):

$$CRPS(F, y) = -\frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} (F(x) - \not{1}\{x \geq y_i\})^2 dx$$

where F is predicted CDF and y_i is observed value.

Other Metrics

Coverage Probability:

$$\text{Coverage} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y_i \in [L_i, U_i]\}$$

Area Under Receiver Operating Curve (AUROC):

- Measures quality of uncertainty estimates for ranking

Visualization

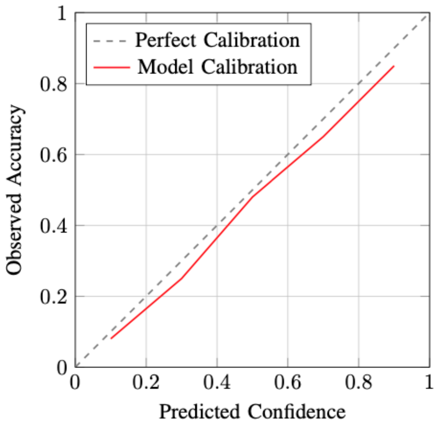

Figure: Calibration plot showing the relationship between predicted confidence and observed accuracy.

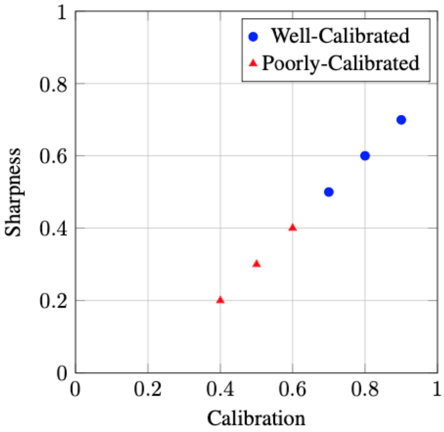

Figure: Trade-off between sharpness and calibration.

Challenges in Uncertainty Quantification (UQ)

- **Computational Complexity & Scalability:**
  Larger, more complex models require *efficient* strategies to handle real-time applications.
- **Interpretability & Usability:** Clear visualization and explanation are key.
- **Disentangling Multiple Uncertainties:**
  Uncertainty becomes complex in multi-modal or large-scale tasks. Proper identification is *crucial* for reliable decisions.
- **Domain-Specific Constraints:**
  Healthcare often involves data privacy; autonomous driving faces real-time constraints.
- **Ethical & Fairness Concerns:**
  Biased uncertainty estimates can lead to unfair decisions (e.g. loan approvals).
- **Lack of Standardization:**
  Absence of unified benchmarks and metrics makes *method comparison* difficult. Common metrics (like calibration error) may not universally apply.

Future Directions in Uncertainty Quantification

- **Advancing Computational Efficiency:**
  Sparse approximations, variational inference, and hardware accelerations (e.g. TPUs) to handle large-scale or real-time UQ.
- **Improving Interpretability:**
  Enhanced uncertainty visualizations and explainable AI (XAI) strategies.
- **Enhanced Uncertainty Modeling:**
  Better disentangling of aleatoric and epistemic uncertainties, especially in multi-modal or temporal data (e.g. deep ensembles, causal inference).
- **Ethical Frameworks for Fair UQ:**
  Fairness-aware methods to mitigate biases and ensure equitable uncertainty estimates.
- **Establishing Benchmarks and Standards:**
  Standardized datasets and metrics to enable reliable evaluation and comparability.

Recommendations

- **Comprehensive Overview**: Offers an extensive summary of recent UQ methods, covering both theoretical foundations and practical tools.
- **Clear, Structured Discussion**: Well-organized sections guide the reader through techniques, metrics, and future directions in UQ.
- **Authoritative References**: Includes up-to-date references and comparisons, helping readers explore the most influential works in the field.
- **Stimulates Further Research**: Highlights open problems and potential research directions, encouraging deeper investigation into novel UQ methods.

References I

Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.