Coverage Testing of MLP Ensemble Classifiers on a 10D Gaussian Toy Model

November 20, 2025

Abstract

We investigate the uncertainty quantification properties of bootstrap-based neural network ensembles for binary classification. Using a 10-dimensional Gaussian toy model with known optimal decision boundary (Neyman-Pearson classifier), we evaluate whether ensemble prediction intervals achieve nominal coverage. Our results show systematic undercoverage, with 95% confidence intervals achieving only 69.9% empirical coverage, indicating that bootstrap ensembles underestimate prediction uncertainty in this setting.

1 Introduction

Uncertainty quantification in machine learning is critical for reliable decision-making, particularly in high-stakes applications. While neural networks can achieve high predictive accuracy, quantifying their uncertainty remains challenging. Ensemble methods offer one approach to uncertainty estimation by training multiple models and using their disagreement as a proxy for uncertainty.

In this work, we systematically evaluate the **coverage** of ensemble-based uncertainty estimates: do the confidence intervals produced by ensembles contain the true optimal predictions at the claimed confidence levels? We use a controlled 10D Gaussian toy problem where the Neyman-Pearson optimal classifier is analytically known, providing ground truth for evaluation.

2 Problem Setup

2.1 Gaussian Toy Model

We consider binary classification with signal ($y = 1$) and background ($y = 0$) classes, each following multivariate Gaussian distributions:

$$p(\mathbf{x}|y = 0) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b) \tag{1}$$

$$p(\mathbf{x}|y = 1) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) \tag{2}$$

where:

- $\mathbf{x} \in \mathbb{R}^{10}$ (10-dimensional feature space)
- Background: $\boldsymbol{\mu}_b = \mathbf{0}$, $\boldsymbol{\Sigma}_b = \mathbf{I}_{10}$
- Signal: $\boldsymbol{\mu}_s = (2, 0, \ldots, 0)^\top$, $\boldsymbol{\Sigma}_s = \mathbf{I}_{10}$

The signal distribution is shifted by 2 standard deviations in the first dimension, providing reasonable class separation.

2.2 Neyman-Pearson Optimal Classifier

The Neyman-Pearson lemma states that the optimal test statistic for binary hypothesis testing is the likelihood ratio:

$$\Lambda(\mathbf{x}) = \frac{p(\mathbf{x}|y=1)}{p(\mathbf{x}|y=0)} \tag{3}$$

For our Gaussian model with equal covariances, this reduces to:

$$\Lambda(\mathbf{x}) = \exp\left(\frac{1}{2}(\|\mathbf{x} - \boldsymbol{\mu}_b\|^2 - \|\mathbf{x} - \boldsymbol{\mu}_s\|^2)\right) \tag{4}$$

The optimal posterior probability (assuming equal priors $p(y=0) = p(y=1) = 0.5$) is:

$$p^*_{\text{NP}}(y=1|\mathbf{x}) = \frac{\Lambda(\mathbf{x})}{1 + \Lambda(\mathbf{x})} \tag{5}$$

This provides the ground truth against which we evaluate our ensembles.

3 Methodology

3.1 Data Generation

- **Training data**: 10,000 samples (5,000 signal, 5,000 background)
- **Test data**: 5,000 samples (2,500 signal, 2,500 background)
- Balanced classes with equal prior probabilities

3.2 MLP Architecture

Each individual classifier is a feedforward neural network with:

- Input layer: 10 units (one per feature)
- Hidden layers: [64, 32] units with ReLU activation
- Dropout: 0.1 probability after each hidden layer
- Output: 1 unit with sigmoid activation (binary classification)
- Loss: Binary cross-entropy
- Optimizer: Adam with learning rate 0.001
- Training: 30 epochs, batch size 512

3.3 Bootstrap Ensemble Construction

To capture both training uncertainty and finite-data uncertainty, we use bootstrap resampling:

1. **Multiple ensembles**: Create $M = 20$ independent ensembles
2. **Models per ensemble**: Each ensemble contains $K = 20$ MLP models
3. **Bootstrap sampling**: Each model is trained on a stratified bootstrap sample of the training data (sampling with replacement while preserving class balance)
4. **Total models**: $M \times K = 400$ models trained

For a given test point $\mathbf{x}$, ensemble j produces predictions:

$$\{p_1^{(j)}(\mathbf{x}), \ldots, p_K^{(j)}(\mathbf{x})\} \tag{6}$$

The ensemble mean and quantiles are:

$$\bar{p}^{(j)}(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^{K} p_k^{(j)}(\mathbf{x}) \tag{7}$$

$$q_\alpha^{(j)}(\mathbf{x}) = \text{quantile}_\alpha \{p_1^{(j)}(\mathbf{x}), \ldots, p_K^{(j)}(\mathbf{x})\} \tag{8}$$

3.4 Coverage Evaluation

Coverage across ensembles measures how often the Neyman-Pearson optimal prediction falls within ensemble confidence intervals:

For confidence level $\alpha \in [0, 1]$, the $(1 - \alpha)$ confidence interval for ensemble j at point $\mathbf{x}$ is:

$$\text{CI}_\alpha^{(j)}(\mathbf{x}) = [q_{\alpha/2}^{(j)}(\mathbf{x}), q_{1-\alpha/2}^{(j)}(\mathbf{x})] \tag{9}$$

For each test point $\mathbf{x}_i$, we compute the fraction of ensembles whose intervals contain the NP prediction:

$$c_i(\alpha) = \frac{1}{M} \sum_{j=1}^{M} \nVdash \left[p_{\text{NP}}^*(\mathbf{x}_i) \in \text{CI}_\alpha^{(j)}(\mathbf{x}_i) \right] \tag{10}$$

The overall coverage at confidence level α is:

$$\text{Coverage}(\alpha) = \frac{1}{N} \sum_{i=1}^{N} c_i(\alpha) \tag{11}$$

Perfect calibration would yield $\text{Coverage}(\alpha) = 1 - \alpha$ for all α.

4 Results

4.1 Individual Ensemble Performance

Table 1 shows metrics for all 20 ensembles at 95% confidence level.

Table 1: Performance metrics for individual ensembles (95% CI)

Ensemble	Coverage	Mean Interval Width	MAE vs NP	RMSE vs NP
1–20 (mean)	0.695	0.0877	0.0259	0.0385
1–20 (std)	0.014	0.0018	0.0005	0.0008

Key observations:

- Individual ensemble coverage ranges from 66.9% to 71.7% (target: 95%)
- Mean absolute error vs NP optimal: ∼0.026 (good prediction accuracy)
- Interval widths are relatively narrow (∼0.088), indicating high confidence

4.2 Coverage Across Ensembles

Figure **??** shows coverage across multiple confidence levels.

Table 2: Coverage across all 20 ensembles at key confidence levels

Confidence Level	Expected Coverage	Empirical Coverage
50%	0.500	0.327
68%	0.680	0.453
90%	0.900	0.633
95%	0.950	0.699
99%	0.990	0.735

4.3 Calibration Analysis

The coverage plot reveals systematic **undercoverage** across all confidence levels:

- At 95% confidence, only 69.9% of predictions are covered
- The gap between expected and empirical coverage is approximately constant (∼25 percentage points)
- This indicates the ensembles are overconfident—their uncertainty estimates are too narrow

5 Discussion

5.1 Sources of Undercoverage

Several factors may contribute to the observed undercoverage:

1. **Bootstrap limitations**: Bootstrap resampling may not generate sufficient diversity among ensemble members. All models see similar data distributions, leading to correlated predictions.
2. **Finite ensemble size**: With only 20 models per ensemble, the quantile estimates may be unstable and systematically biased toward narrower intervals.

3. **Neural network expressiveness**: MLPs with 64-32 hidden units may be overfitting to the training data, resulting in overconfident predictions that don't reflect true uncertainty.
4. **Training procedure**: Reduced training epochs (30) and specific hyperparameters may affect calibration. Models that haven't fully converged might have different uncertainty characteristics.
5. **Lack of explicit calibration**: The raw ensemble predictions are not post-processed for calibration (e.g., via temperature scaling or Platt scaling).

5.2 Potential Improvements

To improve coverage, several approaches could be explored:

- **Increase ensemble size**: Use 50–100 models per ensemble for more stable quantile estimation
- **Deep ensembles**: Train models with different random initializations rather than just bootstrap resampling
- **MC Dropout**: Use Monte Carlo dropout at test time as an alternative uncertainty estimate
- **Temperature scaling**: Apply post-hoc calibration to adjust prediction confidence
- **More training data**: Increase from 10,000 to 50,000+ samples to reduce finite-data effects
- **Conformal prediction**: Use conformal methods to obtain coverage guarantees

5.3 Comparison to Prior Work

Bootstrap ensembles are known to provide well-calibrated uncertainties in some settings (e.g., Breiman's bagging for decision trees), but neural networks present unique challenges:

- Neural networks have high capacity and can memorize training data
- Optimization landscape complexity leads to different local minima
- Bootstrap samples may be too similar for neural networks to explore different hypotheses

Recent work (Lakshminarayanan et al., 2017) suggests that "proper" deep ensembles (with different initializations) achieve better calibration than bootstrap-based approaches.

6 Conclusion

We evaluated the coverage properties of bootstrap MLP ensembles on a 10D Gaussian classification problem with known optimal classifier. Our key findings:

- Bootstrap ensembles exhibit systematic undercoverage across all confidence levels
- 95% confidence intervals achieve only 69.9% empirical coverage
- Ensembles maintain good prediction accuracy (MAE ≈ 0.026) but underestimate uncertainty

- The undercoverage gap is roughly constant (∼25%) across confidence levels

This work highlights the importance of validating uncertainty estimates against ground truth. While ensemble methods provide a practical approach to uncertainty quantification, they require careful evaluation and potentially post-hoc calibration to achieve nominal coverage.

Future work should explore alternative ensemble strategies (deep ensembles, MC dropout) and calibration methods (temperature scaling, conformal prediction) to improve coverage while maintaining computational efficiency.

Code Availability

All code for this project is available at `/global/u1/i/ipang001/NN_UQ/`.

References

[1] Breiman, L. (1996). *Bagging predictors*. Machine Learning, 24(2), 123–140.

[2] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). *Simple and scalable predictive uncertainty estimation using deep ensembles*. Advances in Neural Information Processing Systems, 30.

[3] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). *On calibration of modern neural networks*. International Conference on Machine Learning, 1321–1330.

[4] Neyman, J., & Pearson, E. S. (1933). *On the problem of the most efficient tests of statistical hypotheses*. Philosophical Transactions of the Royal Society of London. Series A, 231, 289–337.