Comparing uncertainty bounds between uncertainty-based methods in deep learning models

Dishant Patel, Hritik Ladia, Utkarsh Mittal
Indian Institute of Technology Gandhinagar

Abstract

This report explores and compares uncertainty bounds in deep learning models, focusing on four prominent methods for predicting intervals: Deep ensemble, MC Dropout, Laplace approximation, MLP with conformal prediction (CP). The primary objectives are to provide an overview of these methods, emphasizing their advantages and disadvantages in uncertainty quantification, delve into the accuracy of prediction intervals, and offer ways to compare uncertainty using metrics such as coverage probability, expected calibration error (ECE), and average width. The study includes an in-depth experimental comparison across four synthetic data sets, revealing insights into variation of performance of different methods, and practical challenges. The results indicate that conformal methods exhibit high reliability, while over-confidence is a common issue across methods. The dependence of uncertainty quantification on the methods used and the data set in question is highlighted.

Introduction

Machine learning's pivotal role in high-impact AI applications, including self-driving cars, medical diagnostics, and machine translation, hinges on the assurance of system reliability. Recent research highlights the criticality of understanding system boundaries, focusing on concepts like ”uncertainty quantification” and ”uncertainty estimation,” particularly in domains like computer vision and natural language processing.

In regression, conventional methods primarily function as point predictors, estimating summary statistics—often the conditional mean—without accounting for prediction confidence. More intricate approaches, however, offer prediction intervals that gauge uncertainty. These intervals, wider when uncertainty is higher, can stem from modeling the conditional distribution using Bayesian or ensemble methods. Alternatively, direct estimation techniques, such as conformal prediction methods, provide prediction intervals without fully modeling the distribution. Comparison among these methods is crucial to discern their efficacy, strengths, and limitations in providing reliable prediction intervals. In this project we tried to bridge that gap, by focusing on the following aspects:

1. We give an overview of four general classes of methods that produce prediction intervals: Deep ensemble, MC Dropout, Laplace approximation, and MLP with conformal prediction (CP). Our exploration emphasizing their potential advantages and disadvantages in uncertainty quantification.
2. We plan to delve into the accuracy of prediction intervals, often termed calibration or validity, and establish its connection with both data and model characteristics. Since having accurately calibrated prediction intervals holds significant importance in numerous applications.
3. We offer ways to compare uncertainty among various models using different metrics such as coverage probability, expected calibration error (ECE), and average width.
4. We provide an in-depth experimental comparison of the four mentioned methods based on their performance across four synthetic data sets. We interpret the observed differences and discuss practical difficulties.

Methodology

Datasets

We make four synthetic datasets with varying generative functions and nature of aleatoric noise. The datasets are as follows:

1. **Linear dataset with Heteroscedastic noise:** This is a linear function with noise increasing with distance from 0. It is given as the below equation in 1 :

$$y = x + \epsilon \tag{1}$$

$$\epsilon \sim \mathcal{N}(0, |X|) \tag{2}$$

2. **Linear Dataset with sinusoidal heteroscedastic noise:** In this dataset we vary the noise in a sinusoidal fashion. We aim to capture the behavior of our techniques when noise is varying but constrained. The dataset is illustrated in 2

$$y = x + \epsilon \tag{3}$$

$$\epsilon \sim \mathcal{N}(0, 1 + sin(X)) \tag{4}$$

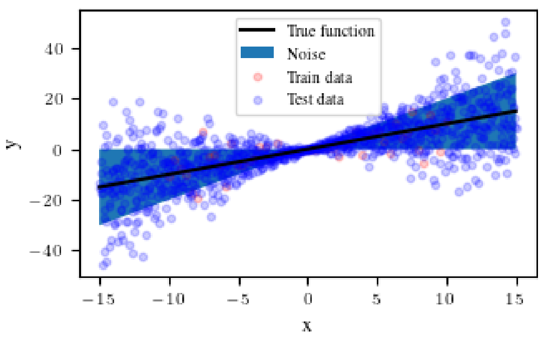


Figure 1: Linear Dataset with heteroscedastic aleatoric noise proportional to distance from 0.

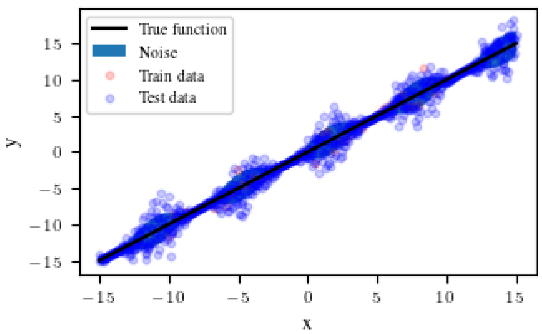


Figure 2: Linear Dataset with sinusoidal heteroscedastic noise. The noise varies but is constrained.

3. **Quadratic dataset with homoskedastic noise:**Keeping noise constant, we aim to study the behavior of functions when the function value blows up. The dataset is illustrated in 3

$$y = x^2 + \epsilon \tag{5}$$

$$\epsilon \sim \mathcal{N}(0, 1) \tag{6}$$

4. **Sinusoidal dataset with homoscedastic uncertainty:** We take sinusoidal dataset to investigate model's behavior upon encoundering more complex data. Data is illustrated in 4

$$y = sin(x) + \epsilon \tag{7}$$

$$\epsilon \sim \mathcal{N}(0, 1) \tag{8}$$

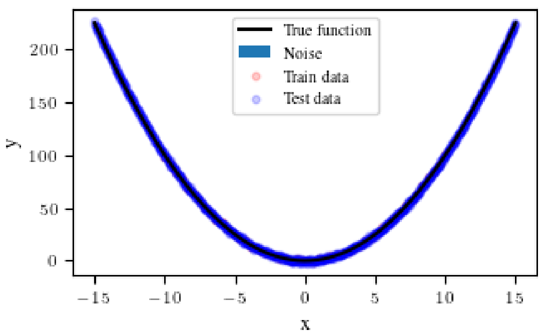


Figure 3: Quadratic dataset with homoscedastic noise. Function blows up at extremes while noise remains constant

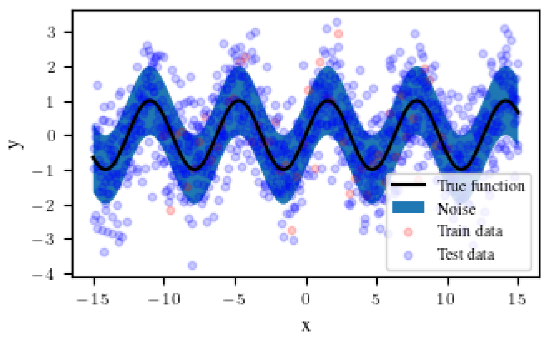
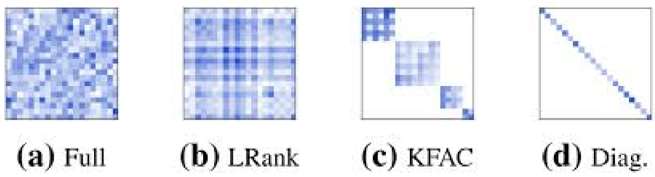


Figure 4: Sinusoidal dataset with constant uncertainty



(a) Full (b) LRank (c) KFAC (d) Diag.

Figure 5: (a)Full Hessian Matrix. (b) Low-rank approximation of hessian. (c) KFAC Approximation. This approximation assumes covariance between parameters of different layers to be 0, hence assuming all layers are independent of one another. (d) Diagonal Approximation. This approximation assumes all parameters are independent to one another.

Methods compared

We use uncertainty-based methods available for Neural networks. These include both Bayesian and Non-Bayesian techniques. The list of methods used is as follows:

1. **Laplace approximation:** We use `laplace-torch` (Daxberger et al. 2021) library for this. We use the last-layer method and compare across all three Hessian approximations: Full Hessian, Diagonal Approximation and Kronecker-Factor Approximate Curvature.

2. **MC Dropout:** We use dropout at last layer, similar to Laplace approximation. We use dropout probability as 0.5. We take 50 predictions for inference.

3. **Deep Ensemble:** We use 5 different models for inference.

4. **Conformal Predictions:** We use the Model Agnostic Prediction Interval Estimator (MAPIE) (Vianney Taquet 2022). (Angelopoulos and Bates 2021) (Dewolf, Baets, and Waegeman 2022). `MLPRegressor()` module is taken from `scikit-learn` library.

Our neural network consists of an input layer of size 1, an output layer of size 1 and 4 hidden layers of size 64 each.

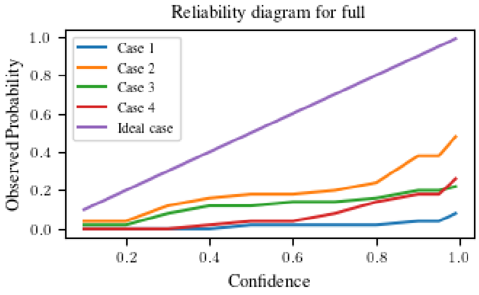


Figure 6: Reliability diagram for Laplace approximation with Full hessian. Reliability is highest for case 2, where noise is constrained.

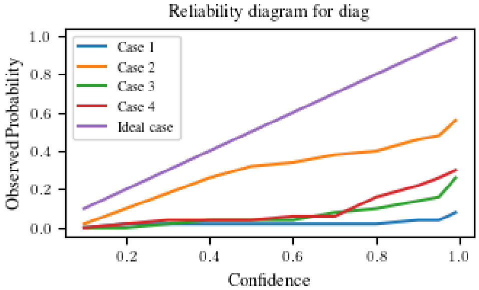


Figure 7: Reliability diagram for Laplace approximation with diagonal Hessian approximation.Reliability is highest for case 2, where noise is constrained.

Metrics

We mainly use metrics pertaining to uncertainty calibration. The metrics we use are

1. **Coverage probability:** This is defined as the probability that a given confidence interval will include a true value. For our case, we report coverage probability for 95% confidence interval.
2. **Expected calibration error:** This is defined for regression as the expectation of the difference between the given confidence interval and the true probability of a point lying in that interval. The mathematical form for it is:

   $$ECE(M) = \mathbf{E}[|c_i - p_i|] \tag{9}$$

   where

   - c_i is the probability for a given confidence interval.
   - p_i is the actual probability of a point for which the probability c_i is output by the model M
3. **Average width**: This is defined as the average width of a given confidence interval for all points in test data. Similar to coverage probability, we capture this quantity for 95% confidence interval.

Results and analysis

The reliability diagrams for different models are given in 6, 8, 7, 9, 10, and 11

As shown in the above results:

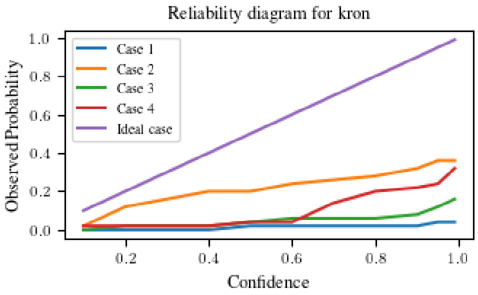


Figure 8: Reliability diagram for Laplace approximation with KFAC approxiimation.Reliability is highest for case 2, where noise is constrained.

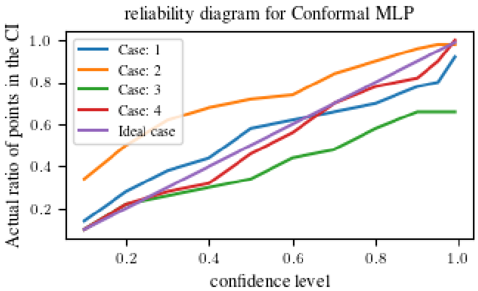


Figure 9: Reliability diagram for Conformal prediction on MLP.Reliability is highest for case 2, where noise is constrained.

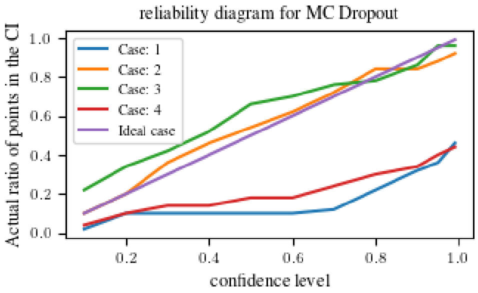


Figure 10: Reliability diagram for MC Dropout.Reliability is highest for case 2, where noise is constrained.

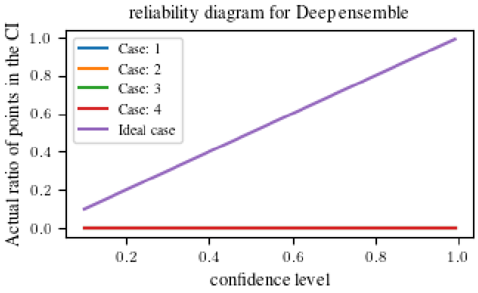


Figure 11: Reliability diagram for Deep ensemble.Reliability is highest for case 2, where noise is constrained.

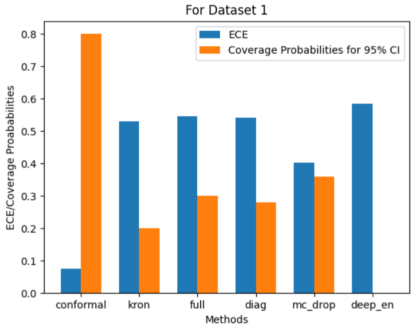


Figure 12: ECE and Coverage for Dataset 1

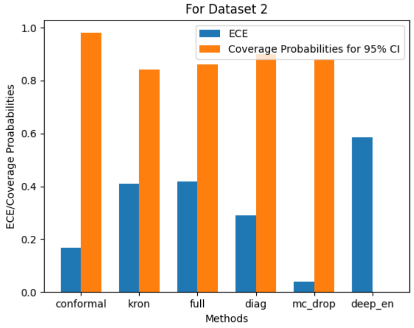


Figure 13: ECE and Coverage for Dataset 2

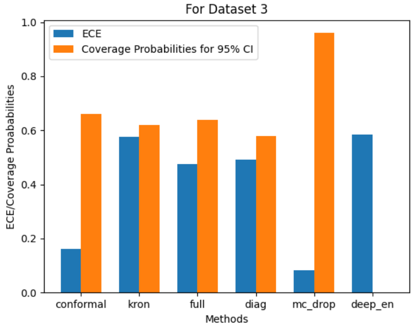


Figure 14: ECE and Coverage for Dataset 3

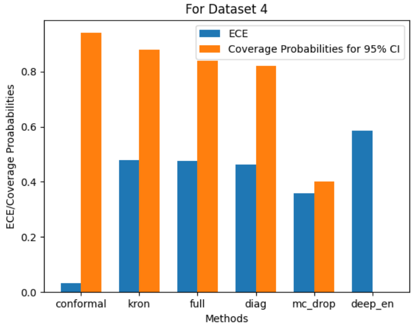


Figure 15: ECE and Coverage for Dataset 4

Method	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Conformal	25.04	5.65	5.37	4.06
Laplace (KFAC)	4.02	3.97	6.03	4.39
Laplace (Full)	4.26	3.95	4.79	4.36
Laplace (Diagonal)	4.10	4.09	4.13	4.17
MC Dropout	6.32	6.87	46.66	1.21
Deep Ensemble	0.29	0.17	9.92	1.73

Table 1: Average width values for each dataset using different methods. Dataset 2 shows a higher width for most datasets.

1. Conformal methods are the most reliable, with the reliability diagram treading closest to the ideal case for all datasets.
2. Except for conformal methods, over-confidence was an issue in almost all methods for all datasets.
3. The third dataset, which had a quadratic function, showed the highest average width. This is because of poor fitting. If the fit is poor, the model must increase the uncertainty bound to accommodate that.
4. Dataset 2 (linear true function with sinusoidal noise) showed the least ECE out of the four. This is because the noise is constrained and does not blow up.
5. Dataset 1 consistently performed poorly on all methods since the noise blows up, meanwhile, the methods assume aleatoric noise. For example, `laplace-torch` library sets a fixed value for aleatoric noise.
6. Deep ensemble performs very badly on all the datasets with very high amount of over-confidence.

Discussion

Over-confidence was an issue in all datasets for almost all methods, except for conformal. This elicits the requirements for new methods to consider the changing uncertainty. We can also reduce uncertainty in Laplace approximation by taking full network instead of last-layer approximation, for which more efficient tools (hardware/algorithmic) for hessian computation and inversion are required. We also need the implementation of hessian approximations for the full network.

The calibration of MAPIE elicits the potential for the library and conformal methods themselves for uncertainty quantification, especially considering the computational challenges for Bayesian methods.

The impact of fitting is elicited by introducing a function for which fit is poor, especially at the extremes. Hence, the quality of fitting of the model highly impacts uncertainty quantification. Hence, uncertainty quantification often depends on the models used and the method in question. The poor performance of deep ensemble on all datasets indicates the requirement for the ensemble size to be more significant. Having a larger ensemble size will more closely resemble the true posterior. Training a high number of deep ensembles, however, requires powerful hardware. Moreover, more significant effort is needed to ensure a greater diversity of models in the ensemble than MC Dropout.

References

Angelopoulos, A. N., and Bates, S. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Daxberger, E.; Kristiadi, A.; Immer, A.; Eschenhagen, R.; Bauer, M.; and Hennig, P. 2021. Laplace redux–effortless Bayesian deep learning. In *NeurIPS*.

Dewolf, N.; Baets, B. D.; and Waegeman, W. 2022. Valid prediction intervals for regression problems. *arXiv preprint arXiv:2107.00363*.

Vianney Taquet, Vincent Blot, T. M. L. L. N. B. 2022. Mapie: an open-source library for distribution-free uncertainty quantification. *arXiv preprint arXiv:2207.12274*.