Сафин, Гнездилова, Малышева, Чернышева

DEEPLY UNCERTAIN: COMPARING METHODS OF UNCERTAINTY QUANTIFICATION IN DEEP LEARNING ALGORITHMS

given enough measurements to characterize the motion of a pendulum, calculate the gravitational acceleration g

Introduction

Motivation and objectives

-> **Motivation**

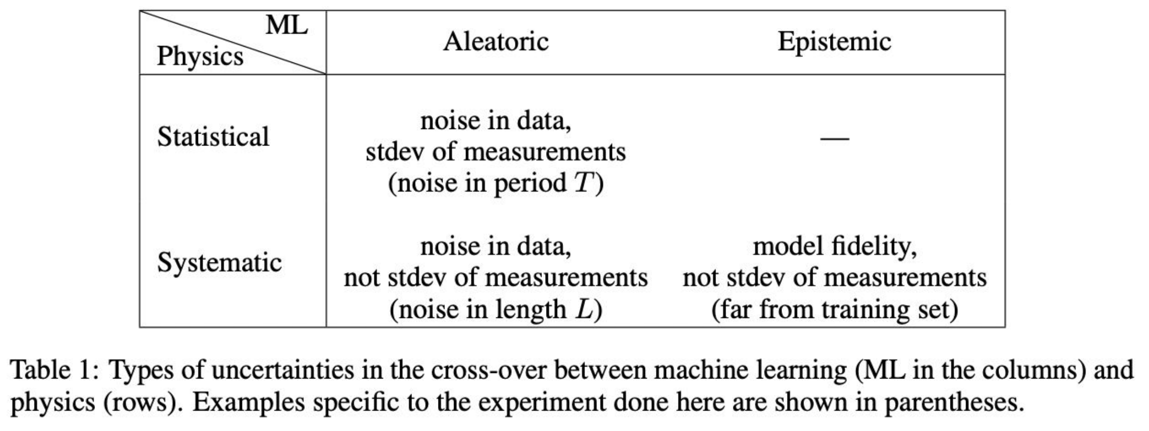
-> Uncertainty quantification (UQ) is crucial for applying deep learning to the physical sciences

-> Different UQ methods have different conceptualizations and interpretations of uncertainty

- **Objectives**
- Compare UQ methods in deep learning for a single pendulum experiment
- Bridge machine learning and physical sciences with a common language and a simple example.
- Evaluate DE, BNN, and CD methods and compare them
- Discuss different types of uncertainty and how they affect data and predictions
- Highlight pitfalls and challenges of UQ methods and make recommendations

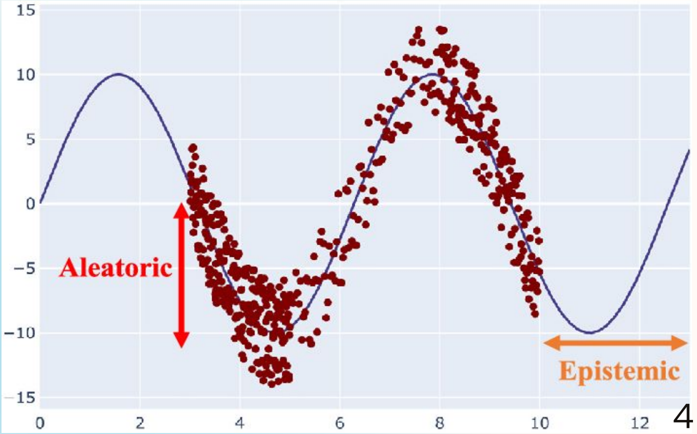
Types of uncertainties

- In machine learning, uncertainties are classified as **aleatoric** or **epistemic**
  a. *Aleatoric uncertainty* originates from noise or variability in the input data
  b. *Epistemic uncertainty* reflects the model's lack of knowledge or confidence in its predictions
- In physics, uncertainties are classified as **statistical** or **systematic**
  a. *Statistical uncertainty* describes the variation of repeated measurements under the same conditions
  b. *Systematic uncertainty* describes the deviation of measurements from the true value due to model assumptions, calibration errors, etc.



ML Physics	Aleatoric	Epistemic
Statistical	noise in data, stdev of measurements (noise in period T)	—
Systematic	noise in data, not stdev of measurements (noise in length L)	model fidelity, not stdev of measurements (far from training set)

Table 1: Types of uncertainties in the cross-over between machine learning (ML in the columns) and physics (rows). Examples specific to the experiment done here are shown in parentheses.



Experiment setup

Inputs: mass *m*, length *L*, angle *θ*, ten independent measurements of the period *T*

$$g = 4\pi^2 \frac{L}{T^2},$$

The output is *g* – gravitational acceleration

Sources of uncertainty

Aleatoric statistical uncertainty

adding noise in the 10 measurements of the period,
T -> vT (v – amount of measurement noise)

Aleatoric systematic uncertainty

all measurements of L are drawn from a normal distribution with standard deviation 0.02L

Epistemic systematic uncertainty

model fidelity, test data is far from the training set

Mathematical formulation of the problem

Inputs: mass *m*, length *L*, angle *θ*, ten independent measurements of the period *T*

$$g = 4\pi^2 \frac{L}{T^2},$$

The output is *g* – gravitational acceleration

We add

The sources of uncertainty are noise

in the measurements of *T* and *L*

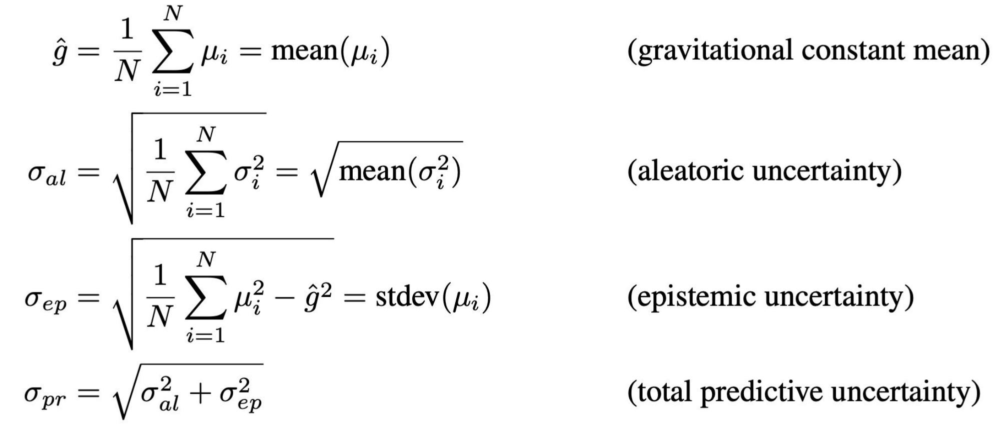
3 methods of UQ: DE, BNN, CD

Calculate statistical uncertainty:

- The spread of predictions between different models is used as an estimate of model-related (i.e., epistemic) uncertainty
- Aleatoric uncertainty is related to the amount of observation noise in a given region of the input space. The effect of that noise on the result is estimated by fitting both the mean and standard deviation of a normal distribution to maximize the log likelihood of the data. The standard deviation obtained is an estimate of aleatoric uncertainty

Mathematical formulation of the problem

For each model and experiment, then, we will obtain N = 10 estimates of the prediction mean and aleatoric uncertainty, (μi, σi). We combine the N estimates as a mixture of Gaussians, and obtain the following predictions:



$$\hat{g} = \frac{1}{N}\sum_{i=1}^{N}\mu_i = \text{mean}(\mu_i) \qquad \text{(gravitational constant mean)}$$

$$\sigma_{al} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\sigma_i^2} = \sqrt{\text{mean}(\sigma_i^2)} \qquad \text{(aleatoric uncertainty)}$$

$$\sigma_{ep} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\mu_i^2 - \hat{g}^2} = \text{stdev}(\mu_i) \qquad \text{(epistemic uncertainty)}$$

$$\sigma_{pr} = \sqrt{\sigma_{al}^2 + \sigma_{ep}^2} \qquad \text{(total predictive uncertainty)}$$

Different UQ methods have different conceptualizations and interpretations of uncertainty

Authors compare three UQ methods: Bayesian Neural Networks (BNN), Concrete Dropout (CD), and Deep Ensembles (DE)

METHODS: EXPERIMENTAL SETUP AND UNCERTAINTY ANALYSIS

Bayesian Neural Networks

Key Features

- weights of each layer form a valid probability distribution
- training via approximate Bayesian inference on probability distributions
- Approximation using the evidence lower bound (ELBO) and Kullback-Leibler (KL) divergence

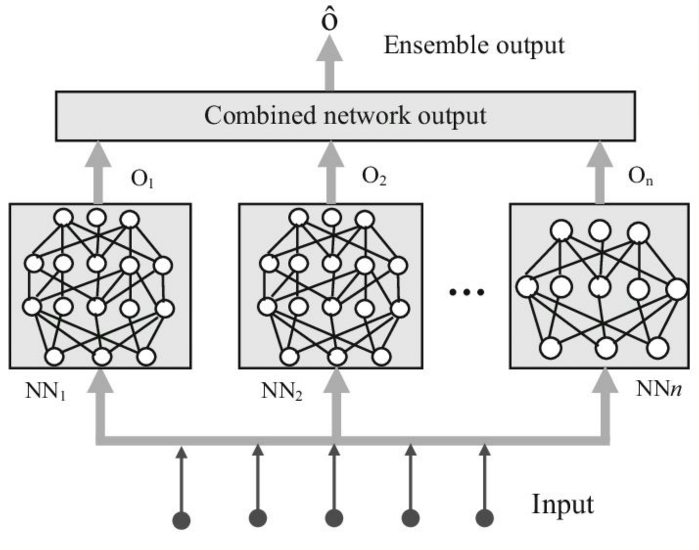
Primary Usage:

Evaluation of epistemic uncertainties by looking at different outputs produced when sampling multiple times from the posterior weight distributions

DEEP ENSEMBLES

Key Features

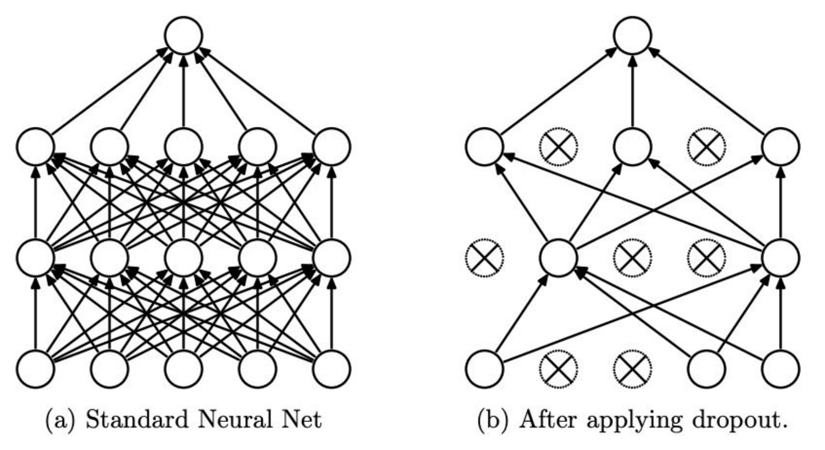
- a simpler alternative to Bayesian methods
- conceptual simplicity
- uses bagging (bootstrap aggregating) for additional randomness



CONCRETE DROPOUT

Key Features

- form of regularization in neural networks
- omitting a certain percentage of neurons at each layer
- drop of a different set of neurons on each pass
- estimates epistemic uncertainties



TRAINING

BNN

- 200 epochs
- learning rate 10^-4

CD

- 200 epochs
- Adam optimizer with
- learning rate 10^-3

DE

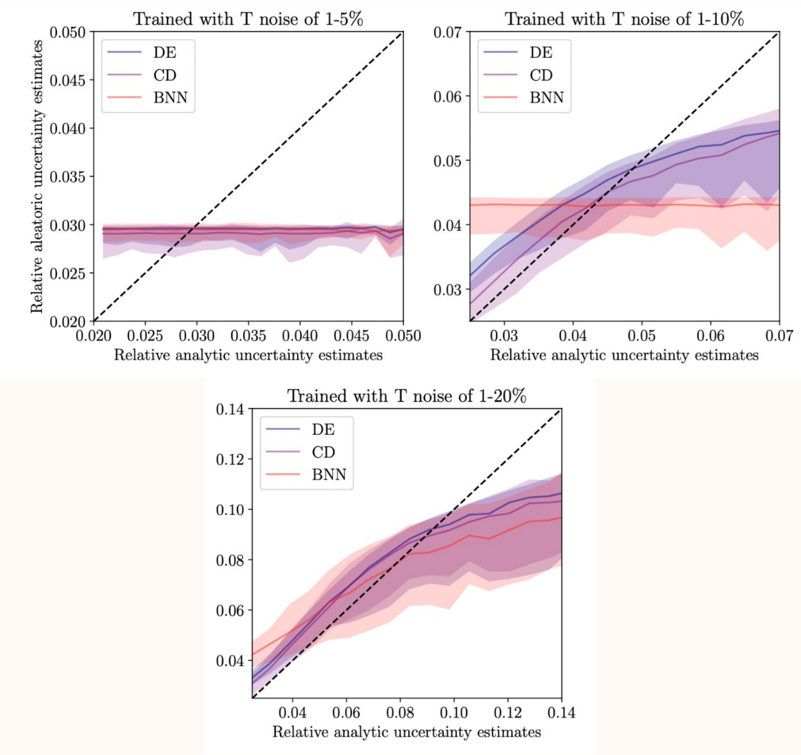
- 40 epochs
- Adam optimizer
- learning rate 10^-3

GENERAL SETTINGS

- fully-connected networks
- 3 hidden layers
- 100 nodes on each hidden layer
- 90000 training points
- TensorFlow 2
- 100 minutes - training time

RESULTS: PERFORMING THE EXPERIMENTS AND UNCERTAINTY ANALYSIS

Aleatoric statistical uncertainty



Comparison of the relative aleatoric statistical uncertainty in g to the relative analytic uncertainty estimate of g for each method, with increasing ranges of noise in T

OUTCOMES

- for larger values of noise in T (middle), DE and CD correlate better with the analytic estimate
- for much larger values of noise in T (right): all methods follow a similar trend to the analytic estimate
- **tendency**: after an initial stage of training, models will often predict the mean value of the training set independently of inputs

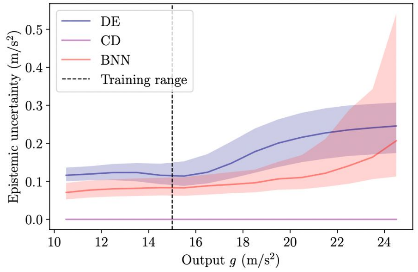
Exploring Epistemic Uncertainties

Experiments with test sets far from the training distribution

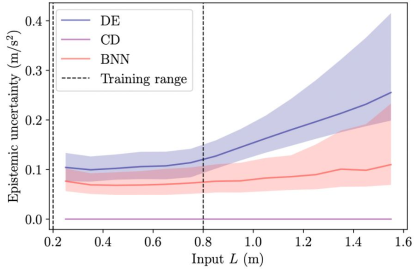
Expectation: the predictions for this uncertainty should increase the farther the input data is from the training distribution

OUTCOMES

- The expected trend is present for both DE and BNN to varying degrees
- CD epistemic uncertainty is very small for a large majority of the points in the test sets presented here, even with increasing distance from the training distribution



(a) Test values of g move away from the network training range, and L remains in range. For the training set, $g \in (5, 15)$ m/s^2.



(b) Test values of L move away from the network training range, and g remains in range. For this training set, $L \in (0.2, 0.8)$ m.

Exploring Epistemic Uncertainties

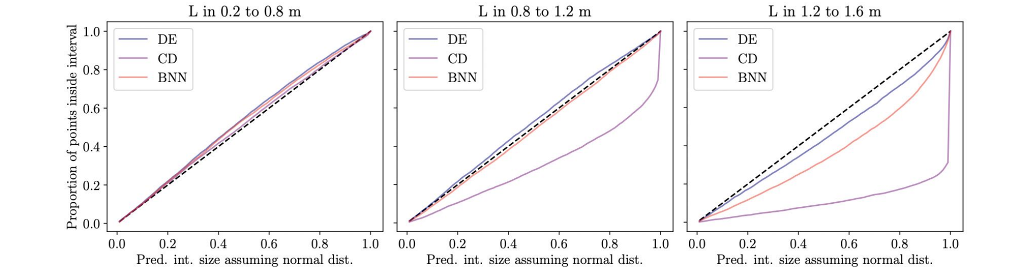


Figure 3: Calibration curves for all UQ methods. The horizontal axis denotes the fraction of the

Testing if the epistemic uncertainties are accurate as the inputs move far from the training manifold, while keeping the outputs inside the training distribution

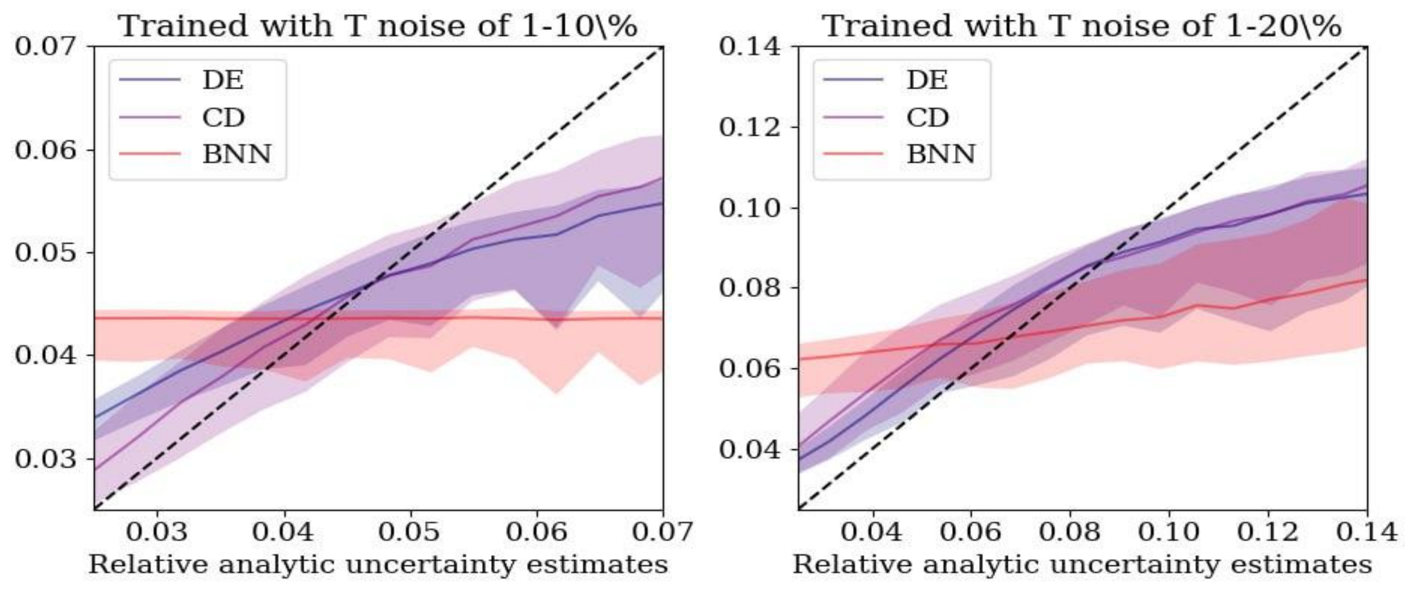
General tendency: they all achieve roughly comparable correlations with the analytic uncertainty estimates, though the correlation achieved by DE is higher

Conclusion and takeaways

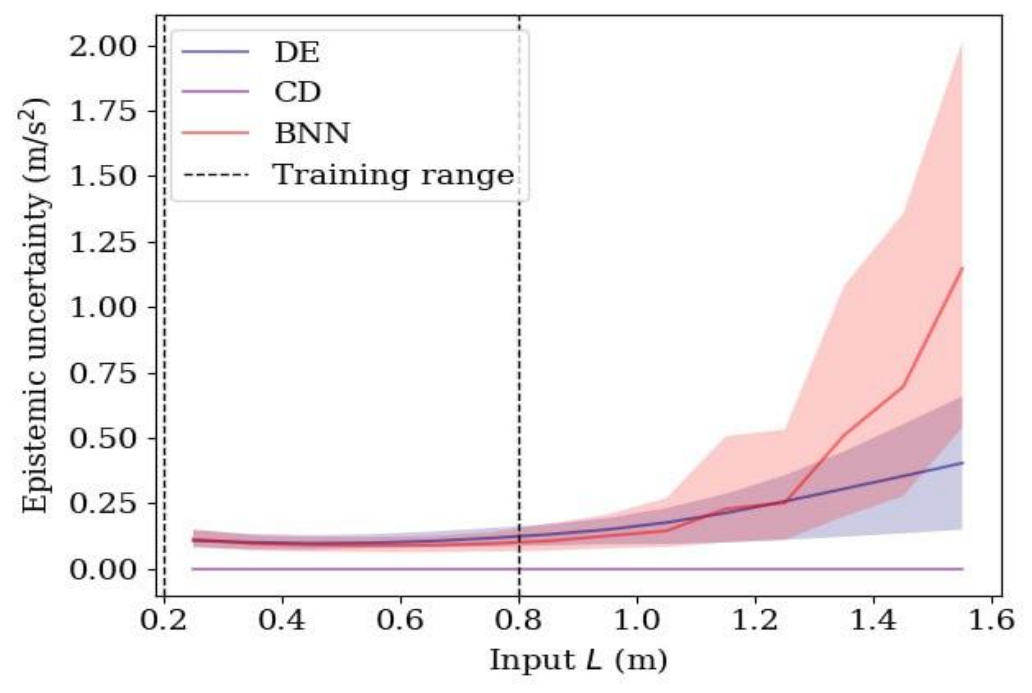
- The authors conclude that UQ is a vital component of deep learning for the physical sciences
- They find that DE is the best method for UQ in this setup, as it has the best performance and simplicity
- They also find that all methods underestimate the epistemic uncertainties when the test data is far from the training distribution
- For aleatoric uncertainties, all methods perform well if the training set has enough variation in noise
- They recommend using reliability diagrams and analytic estimates to assess the calibration and accuracy of UQ methods

Now, to the Notebook!

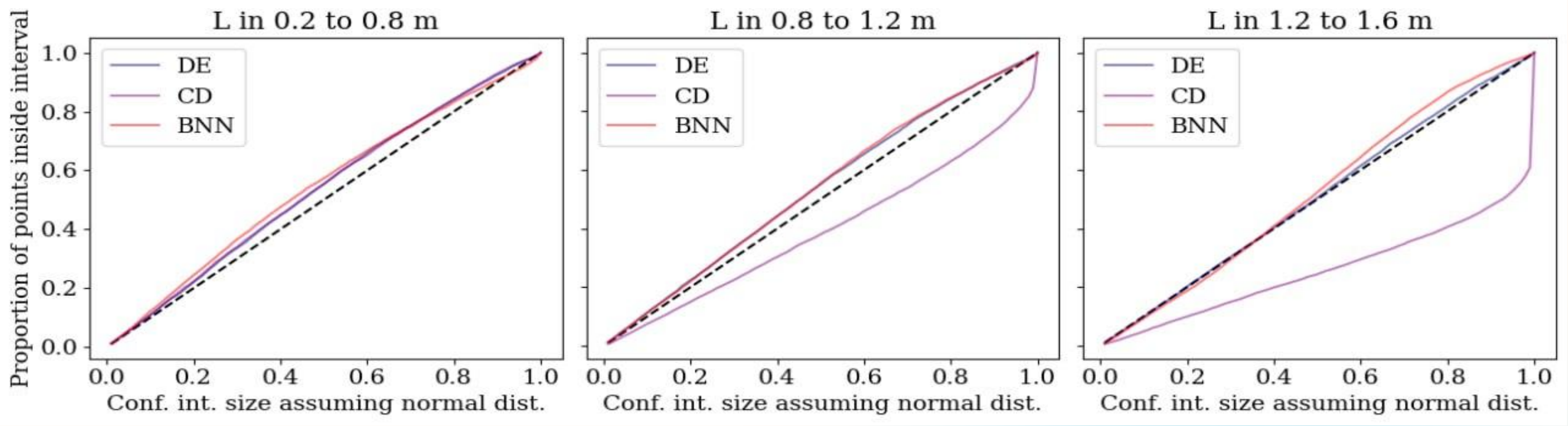
OUR RESULTS


Trained with T noise of 1-10\%
0.07
0.06
0.05
0.04
0.03
DE
CD
BNN
0.03
0.04
0.05
0.06
0.07
Relative analytic uncertainty estimates
Trained with T noise of 1-20\%
0.14
0.12
0.10
0.08
0.06
0.04
DE
CD
BNN
0.04
0.06
0.08
0.10
0.12
0.14
Relative analytic uncertainty estimates

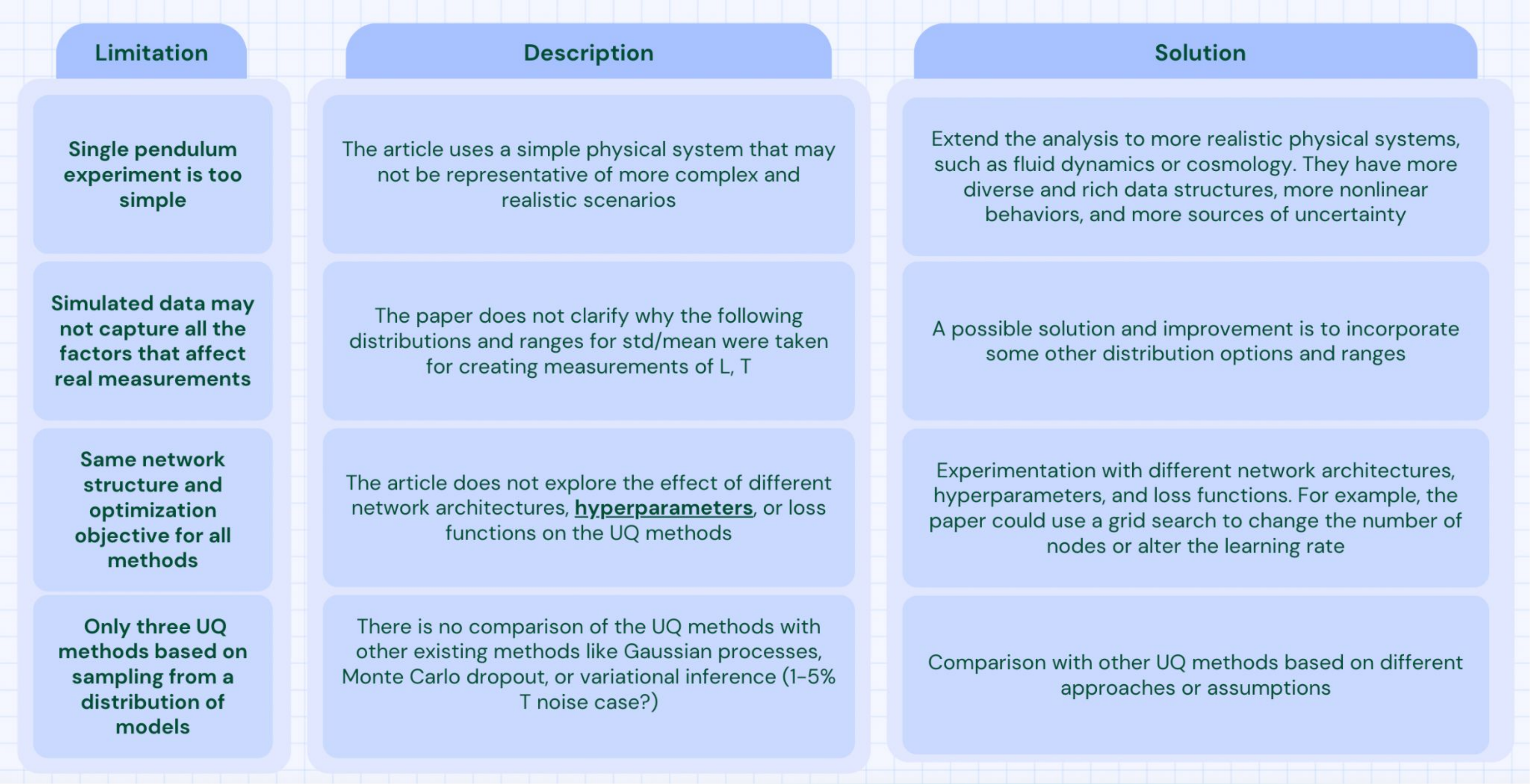
OUR RESULTS


DE
CD
BNN
Training range
Epistemic uncertainty (m/s^2)
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Input L (m)

OUR RESULTS

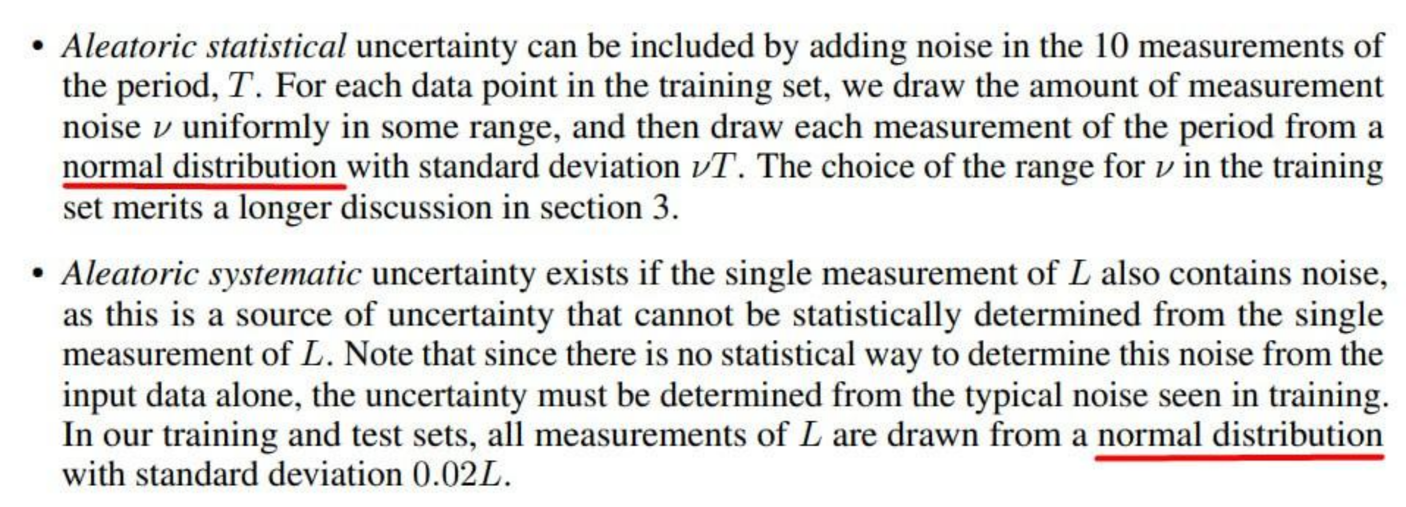

L in 0.2 to 0.8 m
L in 0.8 to 1.2 m
L in 1.2 to 1.6 m
DE
CD
BNN
Proportion of points inside interval
0.0
0.2
0.4
0.6
0.8
1.0
Conf. int. size assuming normal dist.

Limitations and directions

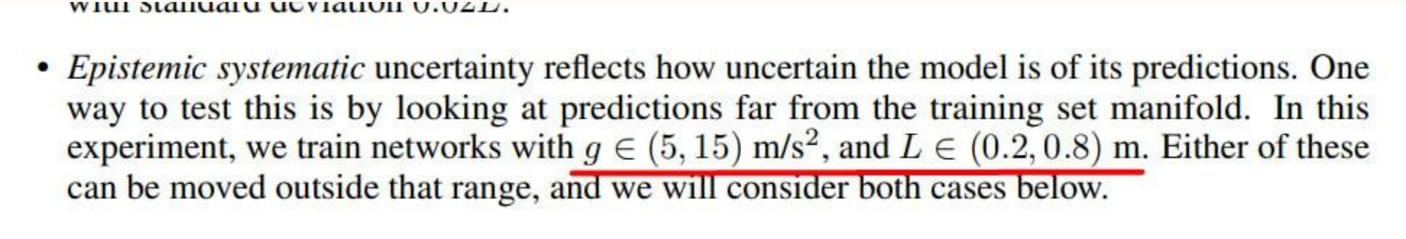


Limitation	Description	Solution
Single pendulum experiment is too simple	The article uses a simple physical system that may not be representative of more complex and realistic scenarios	Extend the analysis to more realistic physical systems, such as fluid dynamics or cosmology. They have more diverse and rich data structures, more nonlinear behaviors, and more sources of uncertainty
Simulated data may not capture all the factors that affect real measurements	The paper does not clarify why the following distributions and ranges for std/mean were taken for creating measurements of L, T	A possible solution and improvement is to incorporate some other distribution options and ranges
Same network structure and optimization objective for all methods	The article does not explore the effect of different network architectures, **hyperparameters**, or loss functions on the UQ methods	Experimentation with different network architectures, hyperparameters, and loss functions. For example, the paper could use a grid search to change the number of nodes or alter the learning rate
Only three UQ methods based on sampling from a distribution of models	There is no comparison of the UQ methods with other existing methods like Gaussian processes, Monte Carlo dropout, or variational inference (1-5% T noise case?)	Comparison with other UQ methods based on different approaches or assumptions

Thank you!



- *Aleatoric statistical* uncertainty can be included by adding noise in the 10 measurements of the period, T. For each data point in the training set, we draw the amount of measurement noise ν uniformly in some range, and then draw each measurement of the period from a normal distribution with standard deviation νT. The choice of the range for ν in the training set merits a longer discussion in section 3.
- *Aleatoric systematic* uncertainty exists if the single measurement of L also contains noise, as this is a source of uncertainty that cannot be statistically determined from the single measurement of L. Note that since there is no statistical way to determine this noise from the input data alone, the uncertainty must be determined from the typical noise seen in training. In our training and test sets, all measurements of L are drawn from a normal distribution with standard deviation $0.02L$.



with standard deviation $0.02L$.

- *Epistemic systematic* uncertainty reflects how uncertain the model is of its predictions. One way to test this is by looking at predictions far from the training set manifold. In this experiment, we train networks with $g \in (5, 15)$ m/s^2, and $L \in (0.2, 0.8)$ m. Either of these can be moved outside that range, and we will consider both cases below.