# On the Effects of Quantisation on Model Uncertainty in Bayesian Neural Networks


Martin Ferianc[1], Partha Maji[2], Matthew Mattina[3], and Miguel Rodrigues[1]

[1]Department of Electronic and Electrical Engineering, University College London, London, UK

[2]Arm ML Research Lab, Cambridge, UK

[3]Arm ML Research Lab, Boston, USA

martin.ferianc.19@ucl.ac.uk, @MartinFerianc, martinferianc, martinferianc.github.io/


UCL

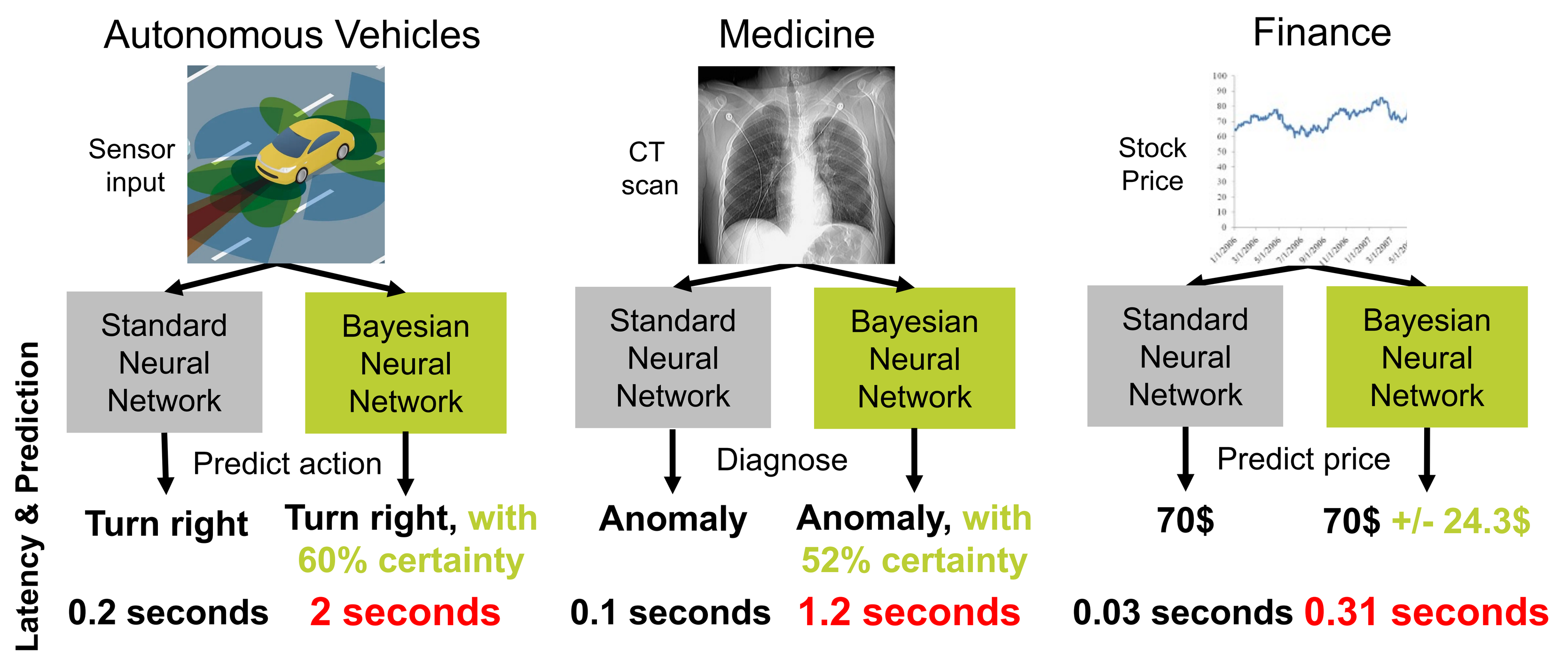

## Summary

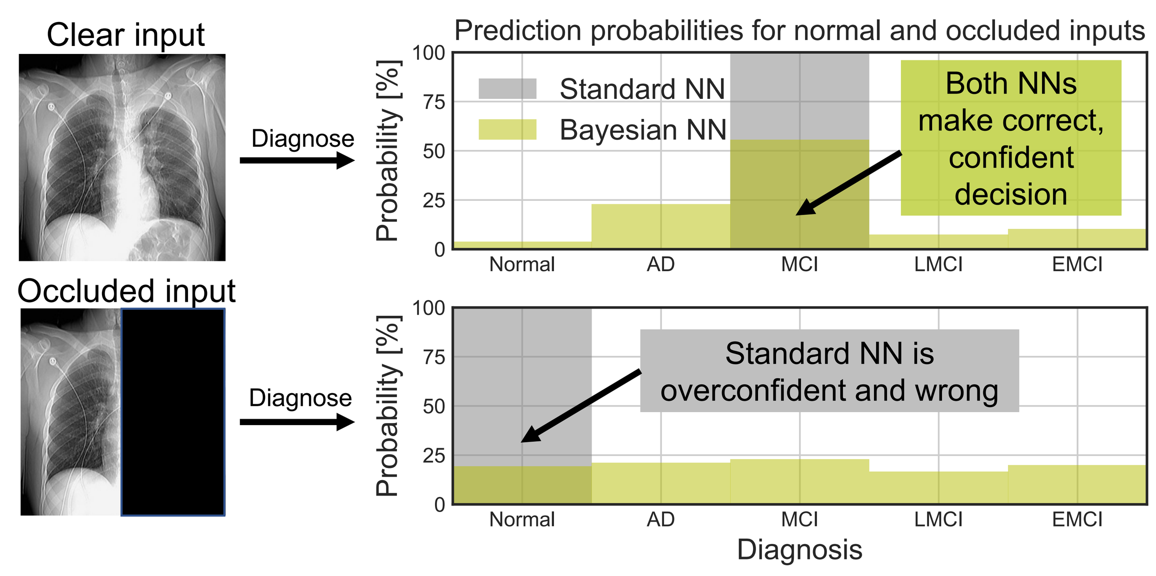
- **What is this about?** Computational representation precision reduction of trustworthy Bayesian neural networks for uncertainty quantification.
- **What is the problem?** Bayesian neural networks, in comparison to standard neural networks, can quantify their uncertainty, but they are $\sim 10\times$ slower.
- **How it is addressed?** Uniform quantisation of Bayesian nets from 32-bit floating point to quantised integers, providing simpler and faster computation.
- **Contribution: Methodology and implementation for quantisation of 3 types of Bayesian neural networks.**


## Introduction

**Uncertainty quantification** (UQ) in machine learning is important for understanding what a model *does not know* and to build trust with users. **Bayesian neural networks** (BNNs) [1] can learn automatically, be accurate and reliably perform UQ.

**Application**: UQ is essential for **safety-critical and regulated real-world applications**, where observing a prediction made by a network is not enough.



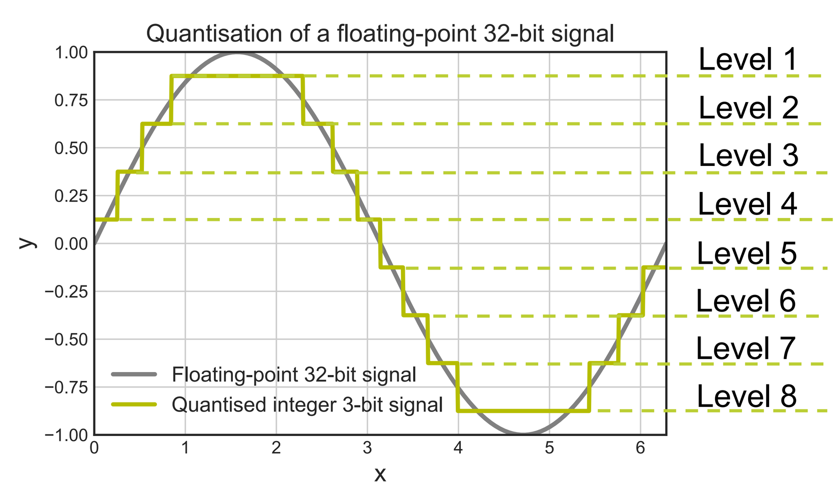




*Example:* network's prediction needs to appear **confused** and **uncertain** by not knowing what to diagnose, then the underlying system can detect an anomaly and a practitioner should have a further look.

**Challenge**: Bayesian neural networks are $\sim 10\times$ **slower**, than standard neural networks with the same architecture and hardware deployment.

**Solution**: Reduce precision of computation through **uniform quantisation** without accuracy or UQ performance loss, generalizable to **most hardware and tasks**.

## Method




*Example:* reduce 32-bit floating-point signal into 3-bit quantised integers.

**Advantages**:

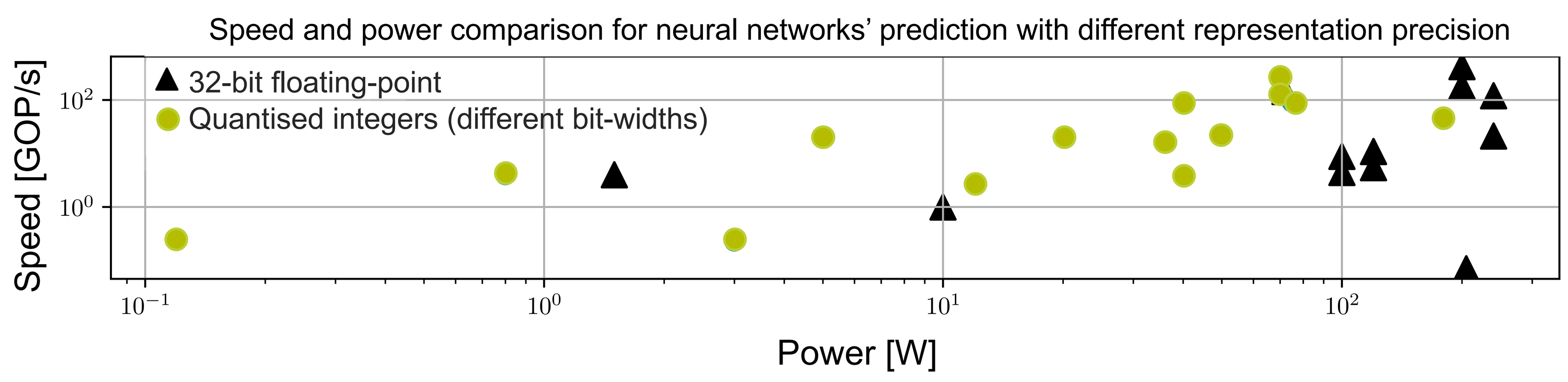
- Represent 8 values (Levels) instead of $3.4\times10^{38}$ values in a computer.
- Generalizable to almost any modern hardware and any application [2].
- Simpler hardware implementation.

$$q = \text{round}\left(\frac{f}{S}\right) + Z$$

Uniform quantisation

- $q$ quantised integer value
- $Z$ zero-point
- $f$ floating-point value
- $S$ quantisation bin-width

$Z, S$ learnable through [2]

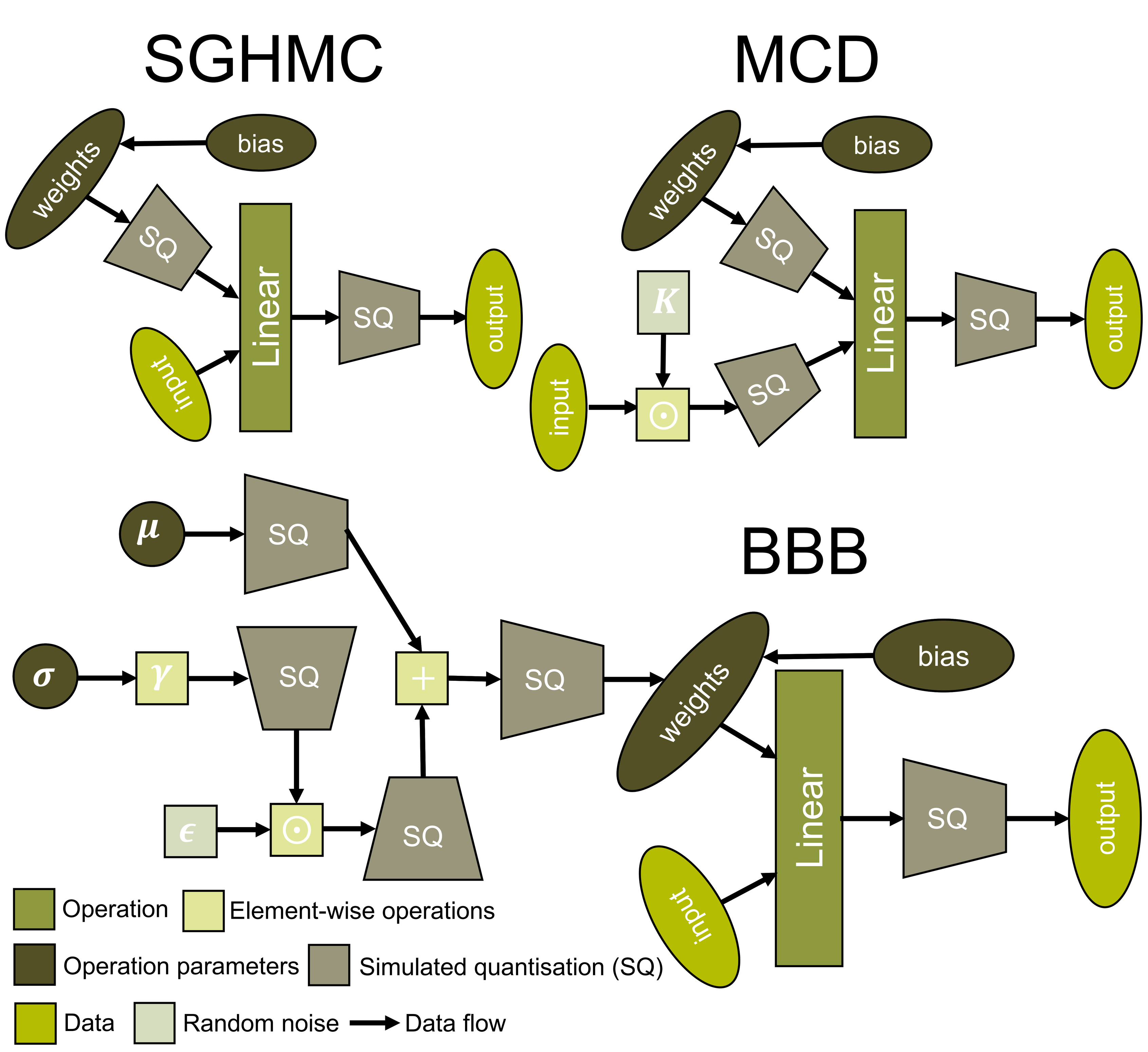



**Quantised computations can be executed faster and cheaper.**

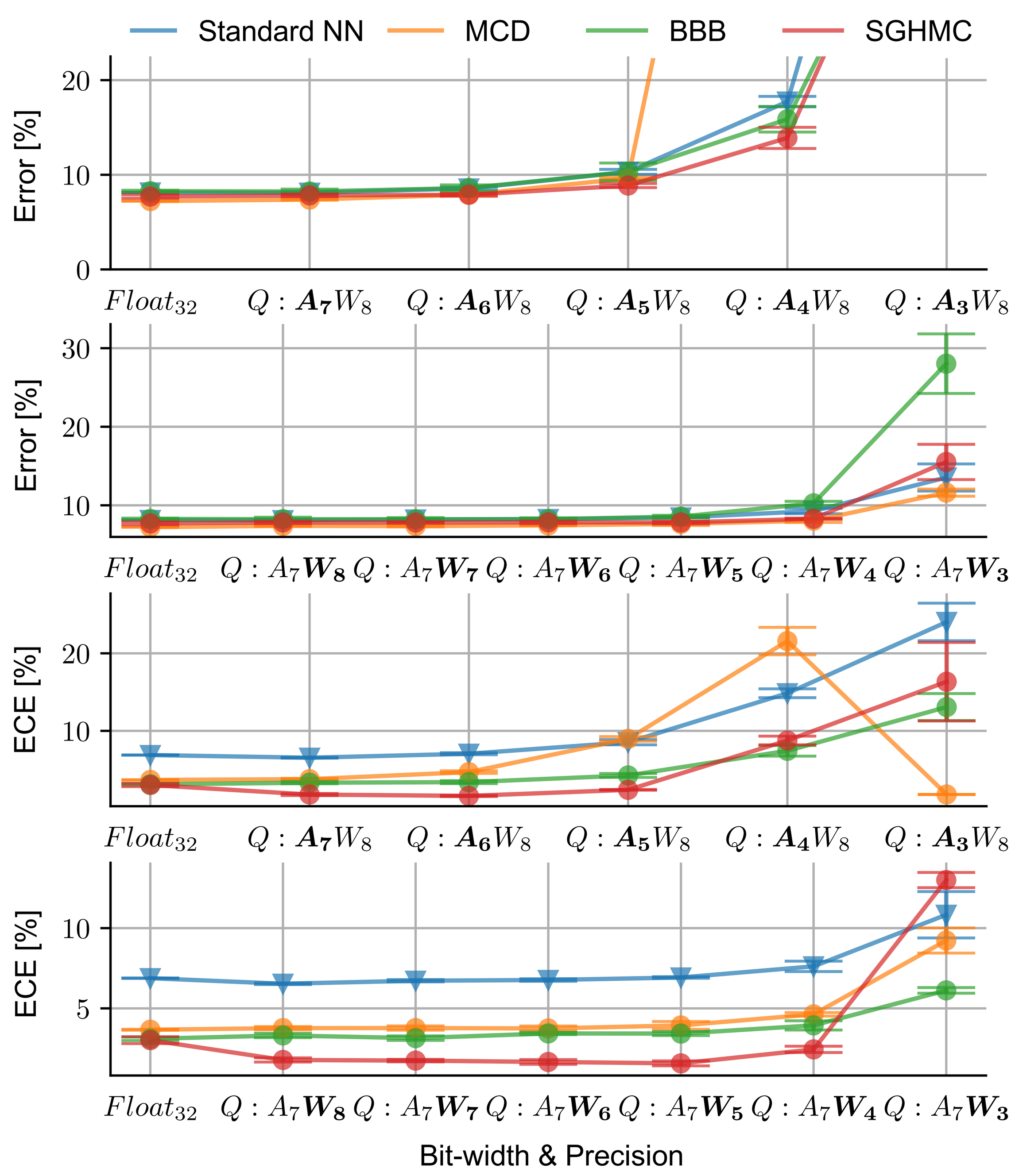
We looked at quantising 3 types of Bayesian neural nets for trade-offs [3]:

- *Bayes-By-Backprop (BBB)*
- *Monte Carlo Dropout (MCD)*
- *Stochastic Gradient Langevin Dynamics with Hamiltonian Monte Carlo (SGHMC)*

**Graphical representation of the proposed quantisation method** for different Bayesian inference schemes shown on a *Linear* layer: output = input·weights + bias.




## Experiments




Experiments with respect to regression (UCI) and classification (MNIST and CIFAR-10) with respect to feed-forward, LeNet and ResNet-18 Bayesian neural network architectures.

Error and expected calibration error (ECE) with respect to quantisation and changing activation (A) precision or weight (W) precision with respect to CIFAR-10 test set. Subscript denotes bit-width.

## Key-Takeaway

**Lowering precision from 32-bit floating-point to quantised 8-bit integers does not detriment accuracy and uncertainty quantification quality of Bayesian neural networks with $\sim 4\times$ compute and memory savings.**

### Room for Improvement and Future Work

Tested Bayesian inference methods were all mean-field approximations which are the least expressive, i.e. could be tried on more complex architectures.

### Paper

### Code


Published at the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence.

**Acknowledgements**: This work was partially completed while Martin Ferianc was an intern at Arm and completed through continued collaboration with Arm ML Research Lab. Martin Ferianc was also sponsored through a scholarship from the Institute of Communications and Connected Systems at UCL.


## References


[1] R. M. Neal, *Bayesian learning for neural networks*, vol. 118. Springer Science & Business Media, 2012.

[2] B. Jacob, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2704–2713, 2018.

[3] H. Wang and D.-Y. Yeung, "A survey on Bayesian deep learning," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–37, 2020.