UNIVERSITY OF CAMBRIDGE

Well-Calibrated Bayesian Neural Networks

Jonathan Heek (jh2098@cam.ac.uk), Tameem Adel (tah48@cam.ac.uk)

The problem

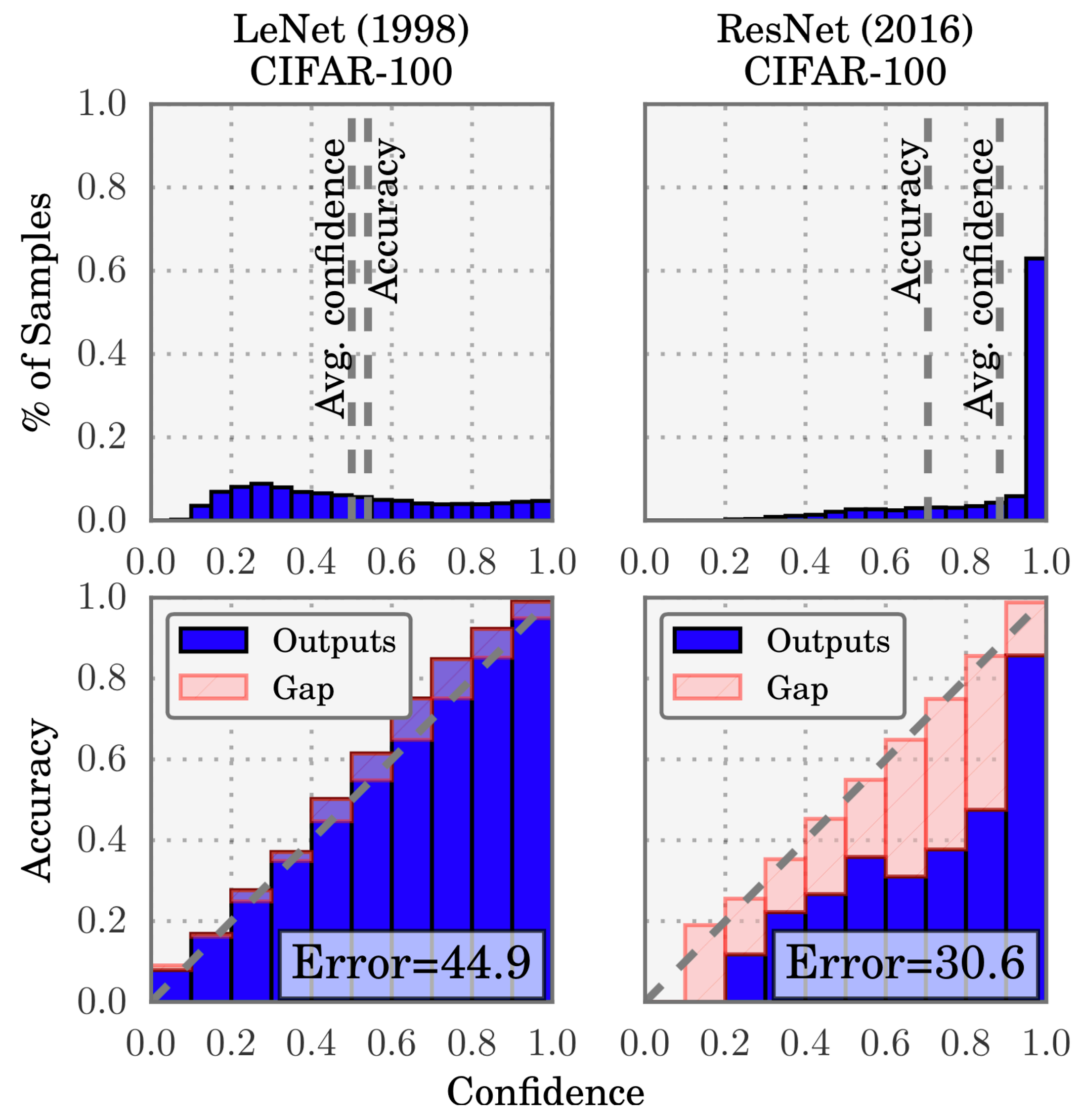

Figure: Prediction confidence vs accuracy of classic and modern neural network, obtained from [1].

Modern neural networks are poorly calibrated meaning the predicted probabilities do not correspond to the observed accuracy.

Measuring Calibration

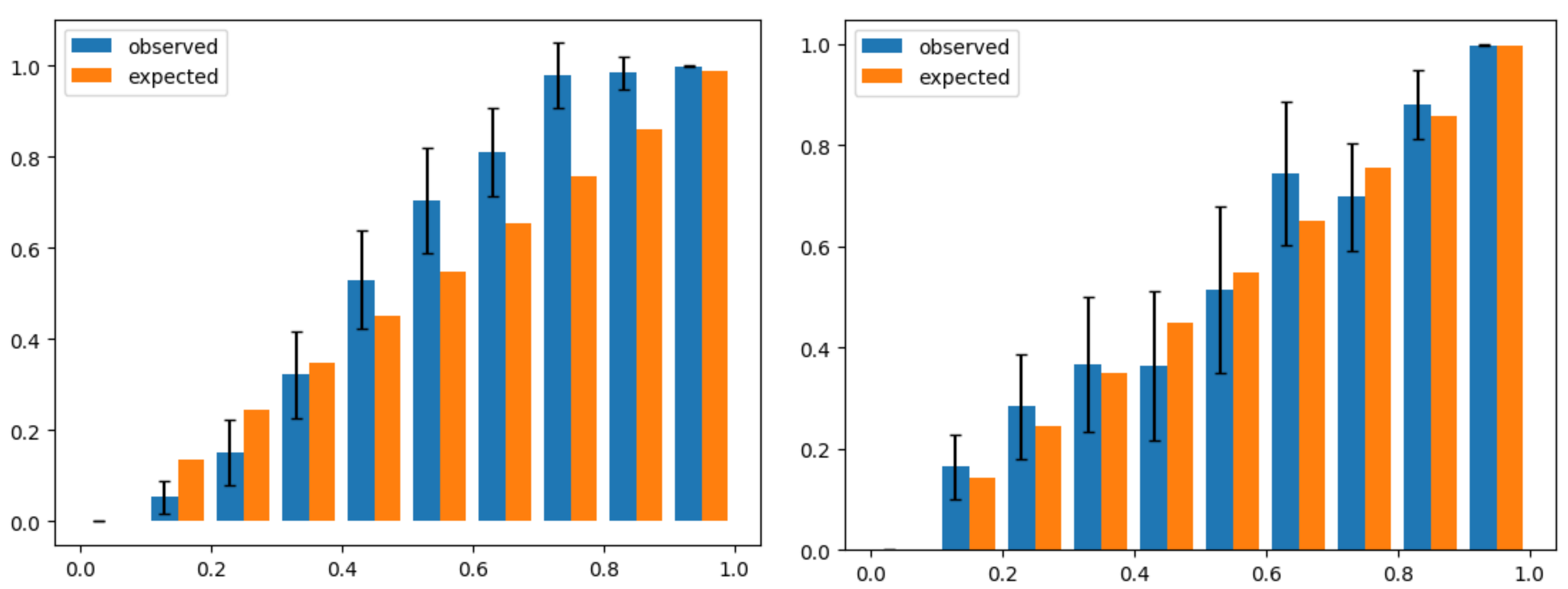

Figure: Calibration plots for Dropout Variational Inference (left) and Bayesian Dark Knowledge (right).

Calibration of a classifier can be measured by binning predictions over confidence levels and comparing the observed and predicted accuracies for the respective confidence levels.

Bayesian Neural Networks

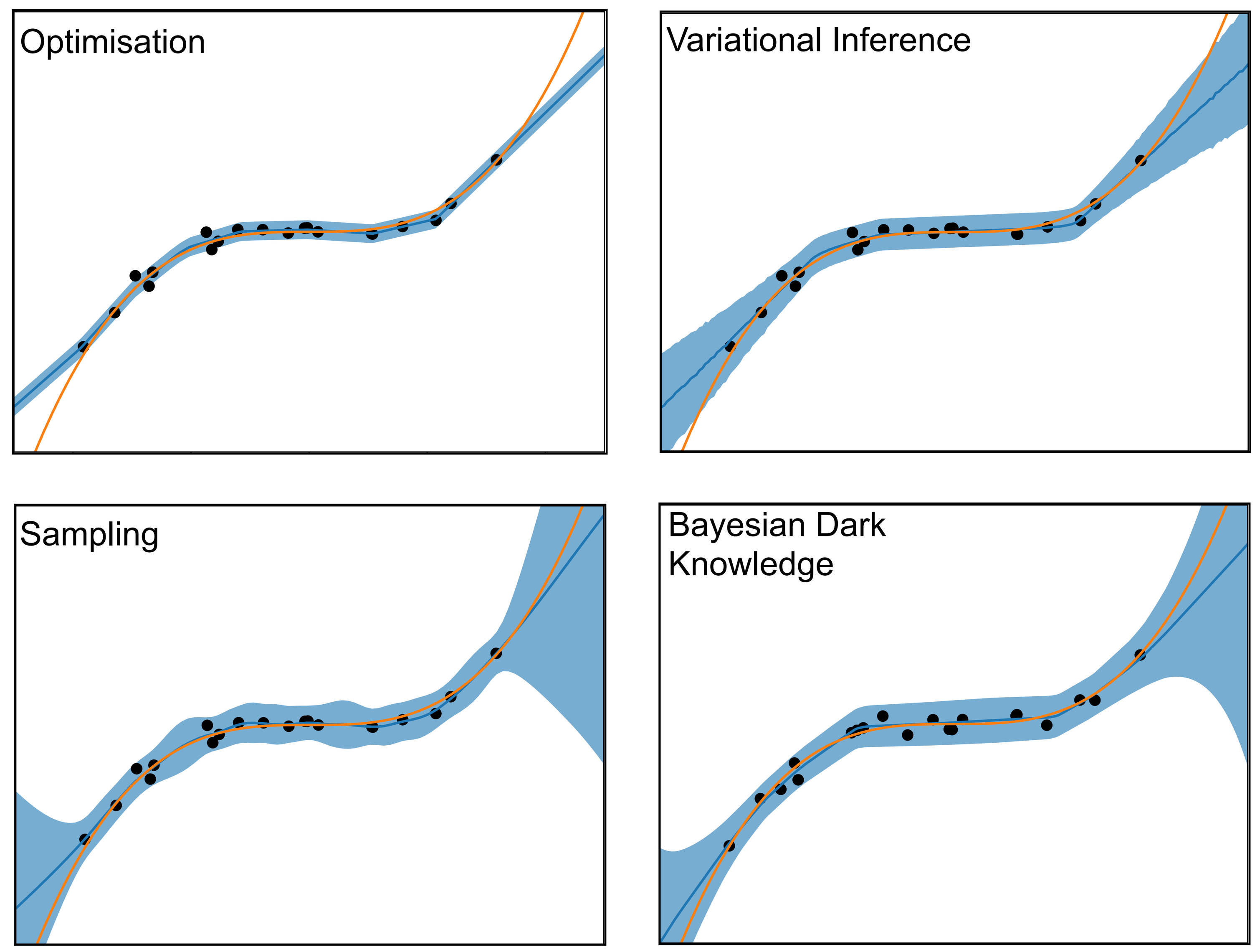

Figure: Uncertainty estimates of a Neural Net with ML inference and and Bayesian Inference methods.

Approximate Bayesian Inference methods significantly improve uncertainty estimates. However, performance and bias are a problem.

Bayesian Dark Knowledge [2] uses student-teacher training to summarise the posterior predictive in a single network which allows constant time inference.

Dropout Variational Inference

$$\mathscr{L}(\theta) = \beta\theta^T\theta - \frac{1}{\alpha}\frac{N}{n}\sum_{i=1}^{n} p(X_i|\theta)^\alpha$$

Dropout Variational inference with **α**-divergences requires only a small modification to the loss function compared to ML inference [3].

Stochastic gradient MCMC

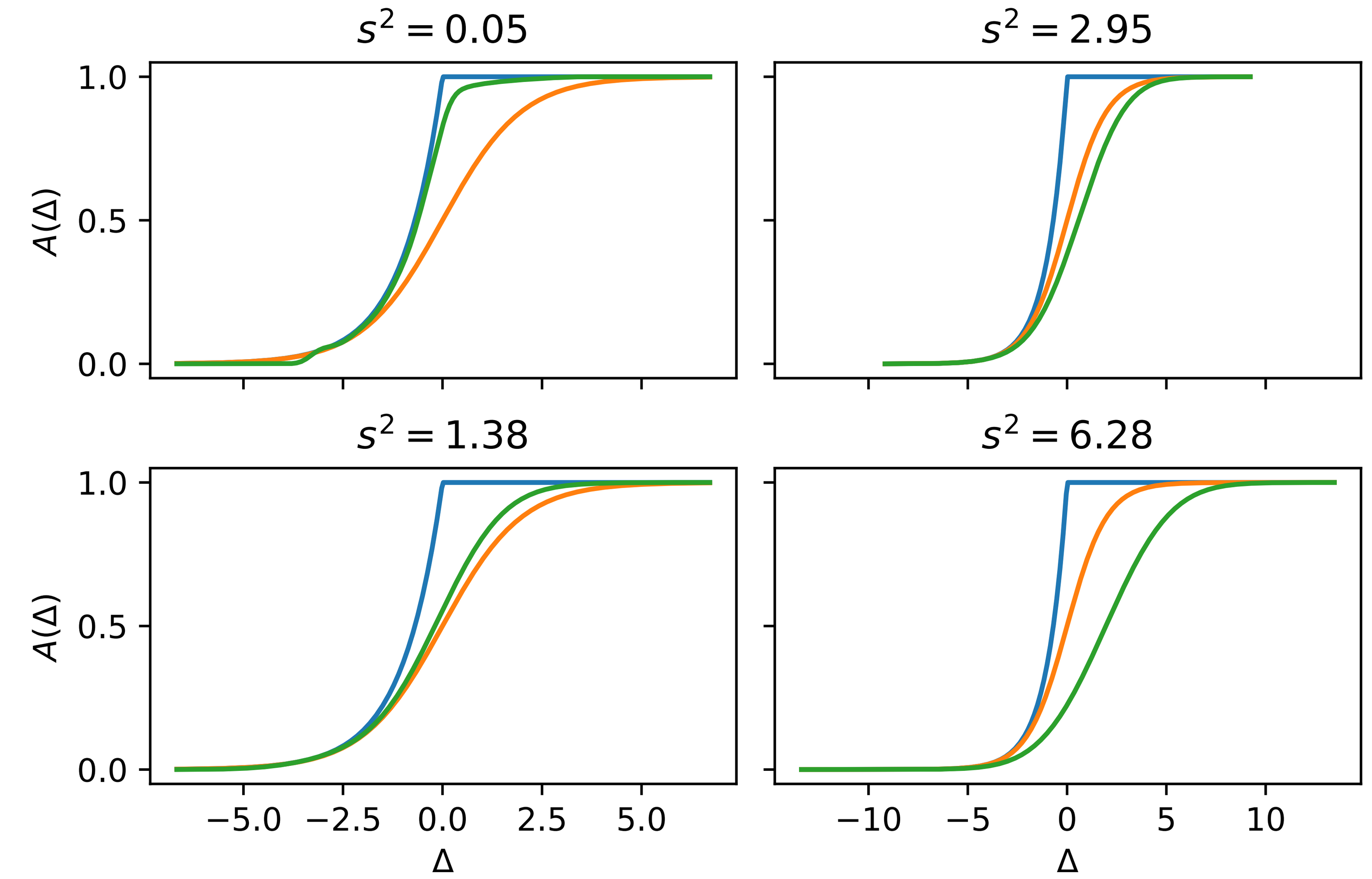

Figure: Metropolis-Hasting (blue), Barker (orange), and noise adaptive acceptance test (green).

The noise adaptive acceptance test is a novel approach to reduce the bias of stochastic gradient HMC.

A tradeoff is made between sample efficiency and acceptance error.

References

[1] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. *ArXiv*.

[2] Korattikara, A., Rathod, V., Murphy, K., & Welling, M. (2015). Bayesian Dark Knowledge.

[3] Li, Y., & Gal, Y. (2017). Dropout Inference in Bayesian Neural Networks with Alpha-divergences. *ArXiv*.