Weight Uncertainty in Neural Networks

John Boom [1] Emma Prévot [1] Ilaria Sartori [1]

[1]University of Cambridge, Department of Engineering

UNIVERSITY OF CAMBRIDGE

Bayesian Neural Networks (BNNs)

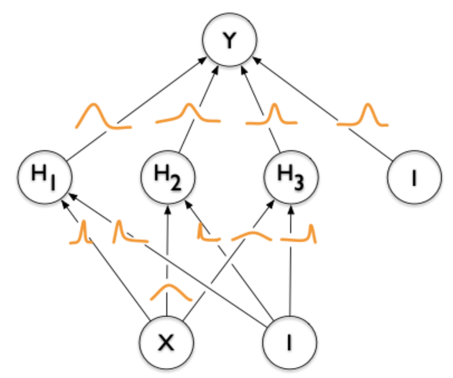

Figure 1. Bayesian Neural Networks.

Represent weights by probability distributions over possible values, rather than a single fixed value

Variational approach: Approximate the posterior $P(w|\mathcal{D})$ with the variational distribution $q(w;\theta)$ minimizing the Kullback-Leibler (KL) divergence

$$\theta^* = \arg\min_\theta \mathrm{KL}\left[q(w;\theta)||P(w|\mathcal{D})\right] = \arg\min_\theta \mathcal{F}(\mathcal{D},\theta)$$

where $\mathcal{F}(\mathcal{D},\theta)$ is called variational free energy

$$\mathcal{F}(\mathcal{D},\theta) = \underbrace{\mathrm{KL}\left[q(w;\theta)||P(w)\right]}_{\text{Complexity cost}} \underbrace{-\mathbb{E}_{q(w;\theta)}\left[\log P(\mathcal{D}|w)\right]}_{\text{Likelihood cost}}$$

Advantages
- Uncertainty estimation
- Regularization

Disadvantages
- Long training time
- Intractable posteriors

Bayes By Backprop (BBB)

Approximate $\mathcal{F}(\mathcal{D},\theta)$ using Monte Carlo:

$$\mathcal{F}(\mathcal{D},\theta) \approx \sum_{i=1}^{n} \log q(w^{(i)};\theta) - \log P(w^{(i)}) - \log P(\mathcal{D}|w^{(i)})$$

where $w^{(i)}$ is the i^{th} MC sample drawn from the variational posterior $q(w^{(i)};\theta)$

Advantages
- Accurate predictions from cheap model averaging

Disadvantages
- Requires MC variance control
- Requires careful prior elicitation

Deterministic Variational Inference (DVI)

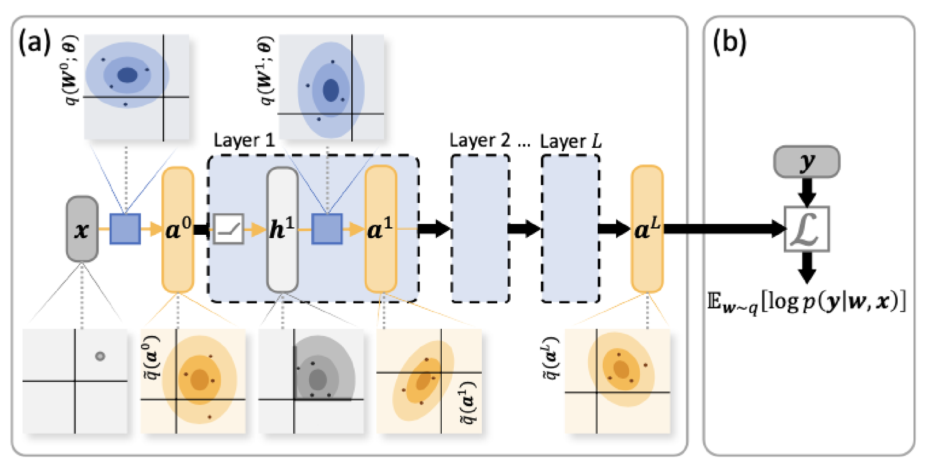

Figure 2. BNN likelihood cost computation.

Likelihood cost:

(a) Activation propagation. Deterministic form to approximate the final layer activation distribution $q(a^L)$

(b) Log-likelihood computation

Complexity cost:

- Closed-form expression for KD
- Hierarchical priors. Empirical Bayes for automatic selection

Advantages
- Remove MC stochasticity
- Automatic prior selection

Disadvantages
- Closed-form limits design
- High compute cost on wide nets

MNIST - Classification

- **BNNs** achieve superior performance compared to **regular FCNs**, with or without dropout, and converge around similar epochs if not earlier. **DVI** achieves comparable performance in fewer (but longer) epochs

		SGD	SGD Dropout	Mixture BBB	Gaussian BBB	DVI
# Weights	480k	97.96	98.22	**98.42**	98.39	-
	2.4m	98.03	98.48	98.50	**98.51**	-
	240k	-	-	-	-	98.02

Table 1. MNIST Classification Accuracy. SGD and BBB methods were trained for 300 epochs, with 400 hidden units (480k) and 1200 hidden units (2.4m). DVI trained only for 30 epochs for computational complexity.

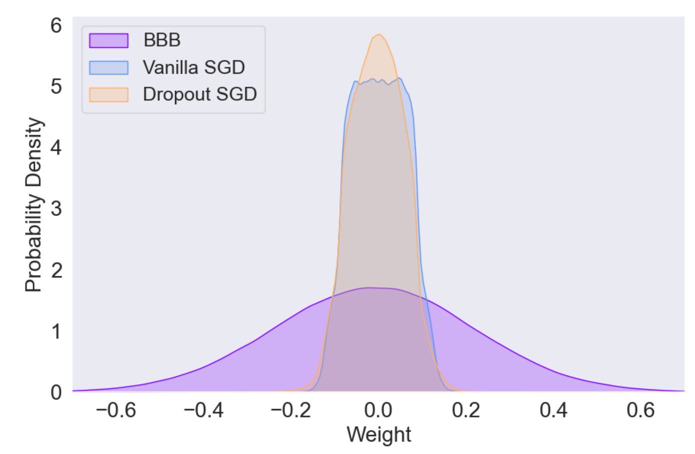

Figure 3. Histogram of the trained weights.

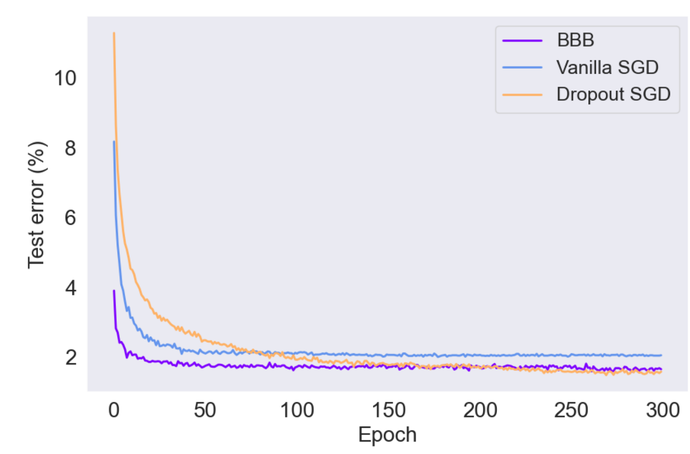

Figure 4. Test error as training progresses.

MNIST Model Weight Pruning

Percentage	0	5	25	50	75	95	99	99.9
# Parameters	480k	460k	360k	240k	120	24k	5k	500
Accuracy (%)	97.2	97.4	97.2	97.3	97.3	97.3	97.1	37.8

Table 2. MNIST classification accuracy after weight pruning of the 400 hidden units Mixture BBB model.

- Carefully choosing the BNN **prior distribution** as well as the **weight initialisation** allows to prune a surprisingly significant percentage of low SNR weights with almost no impact on performance

DermaMNIST - Classification

- Bayesian approaches are well suited for applications where knowing the uncertainty of one prediction is essential, such as Medicine

Method	BBB (400)	BBB (1200)	ResNet-18	Google AutoML Vision
Accuracy	74.9	74.5	73.5	76.8

Table 3. BBB DermaMNIST Classification accuracy against state-of-the-art.

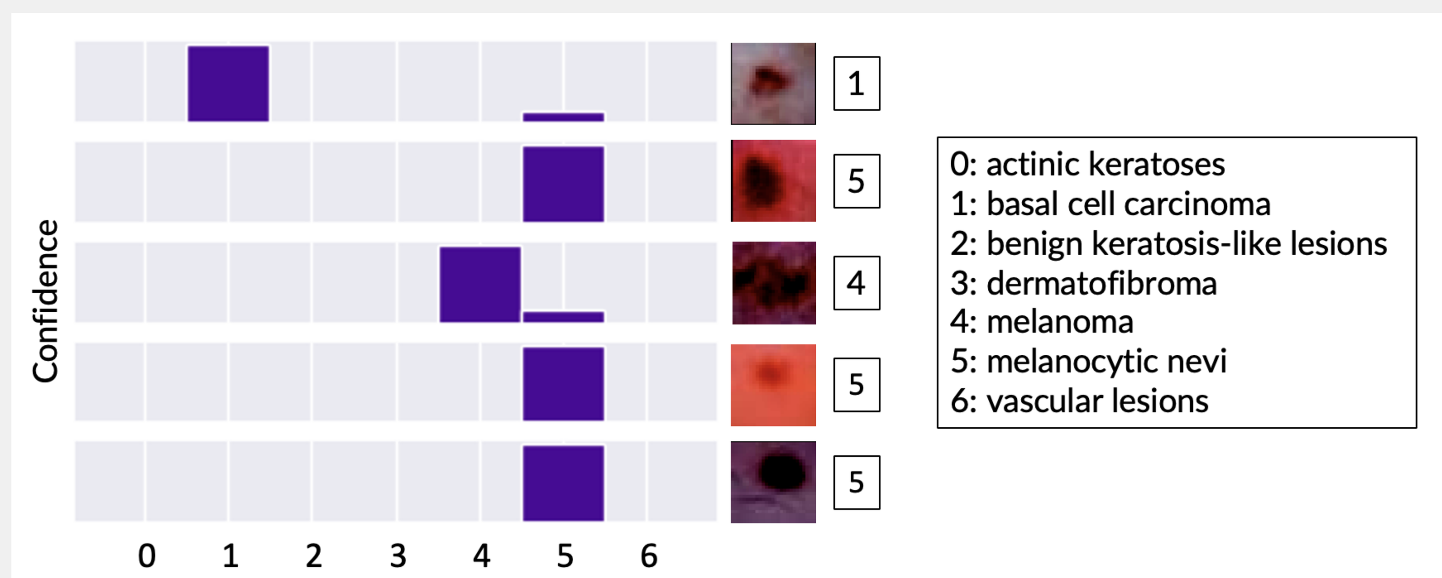

Figure 5. Model diagnosis confidence on dermatoscope pictures from DermaMNIST.

Regression

- Compared to standard NN, a Bayesian approach to Regression allows to obtain **uncertainty estimation** and reduces the risk of overfitting
- DVI with automatic prior selection is slower to converge, but better captures prediction uncertainty, especially in the heteroskedastic and discontinuous datasets

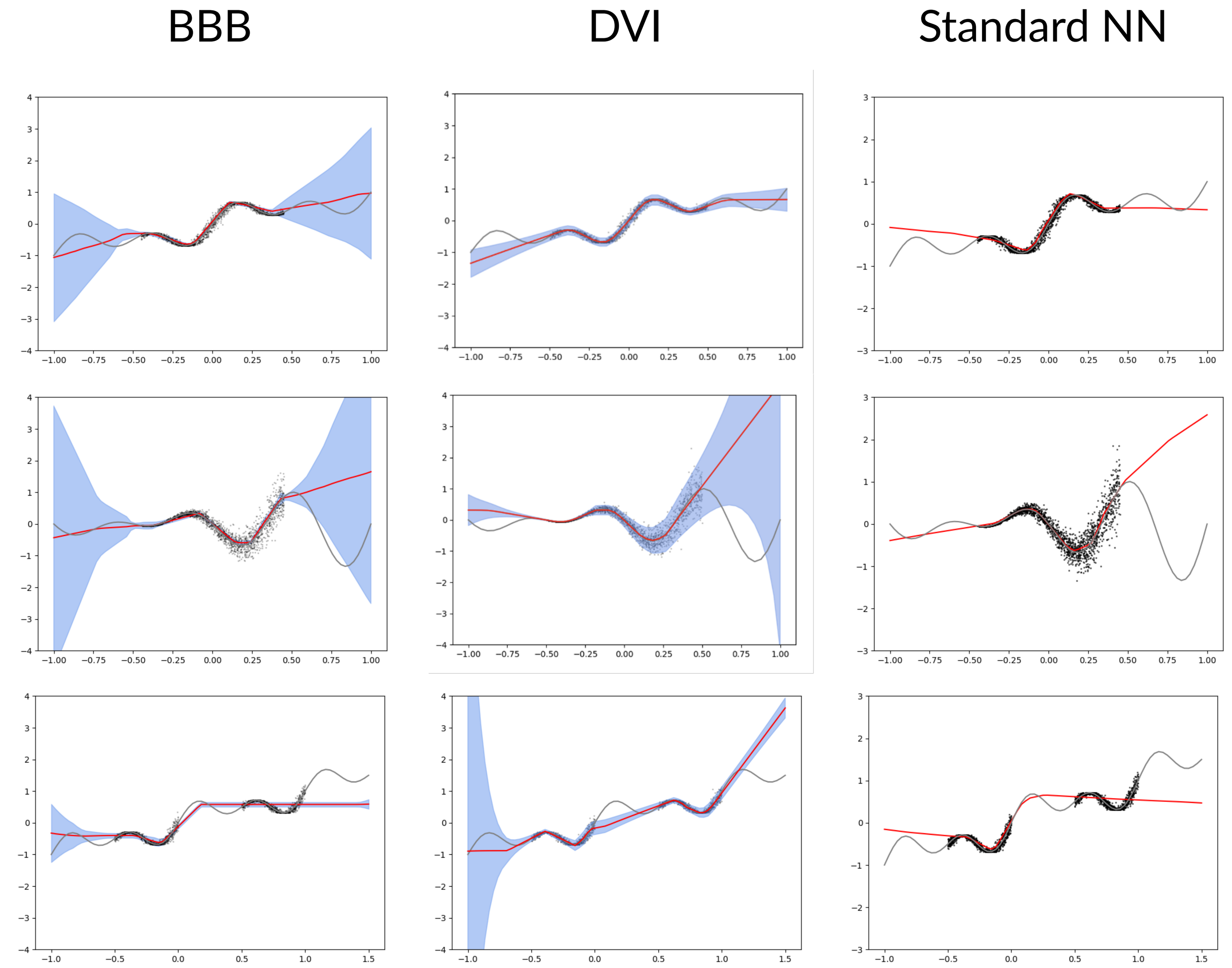

Figure 6. Comparison of BBB, DVI, and a Standard Neural Net on toy datasets with homoskedastic noise (row 1), heteroskedastic noise (row 2), and discontinuous data (row 3). The scattered data-points are the training data, Grey is the true function; Red is the mean prediction; Blue is ±2 standard deviations.

Prior distributions

BBB: $w_j \sim \pi\mathcal{N}\left(0,\sigma_1^2\right) + (1-\pi)\mathcal{N}\left(0,\sigma_2^2\right)$, $\pi \in [0,1]$

DVI: $w_j^\lambda \sim \mathcal{N}(0, s_\lambda)$, $s_\lambda \sim \text{Inv-Gamma}(\alpha,\beta)$

Conclusions and Future work

Key findings
- BNNs are **powerful and incorporate uncertainty**, but fragile to train
- Some models can be **heavily pruned** using SNR
- BBB's **high sensitivity** to weight initialisation and choice of prior yields significant variability

Future work
- Perform **weight pruning on DVI** models
- Apply the **Local Reparametization Trick** (LRT) to Classification tasks with BBB
- Compare BBB and DVI on **Bandit Tasks** where the model can ask for specific new data

References

[1] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," 2015.

[2] A. Wu, S. Nowozin, E. Meeds, R. E. Turner, J. M. Hernández-Lobato, and A. L. Gaunt, "Deterministic variational inference for robust bayesian neural networks," 2018.