Natural gradient descent for improving variational inference based classification of radio galaxies

Devina Mohan
Department of Physics & Astronomy
University of Manchester, UK
devina.mohan@postgrad.manchester.ac.uk

Anna Scaife[*]
Department of Physics & Astronomy
University of Manchester, UK
anna.scaife@manchester.ac.uk

Abstract

Bayesian neural networks (BNNs) are most commonly optimised with first-order optimisers such as stochastic gradient descent. However, when optimising for parameters of probabilistic models, incorporating second order information during optimisation can lead to a more direct path in the distribution space and faster convergence. In this work we examine whether using natural gradient descent can improve the performance of variational inference based classification of radio galaxies. We use the Improved Variational Online Newton (iVON) algorithm and compare its performance against a recent benchmark for BNNs for radio galaxy classification. We find that iVON results in better uncertainty calibration out of all the methods previously considered while providing similar predictive performance to the best performing inference methods such as Hamiltonian Monte Carlo and Bayes by Backprop based variational inference. Models trained with iVON can distinguish far out-of-distribution optical galaxy data, but they cannot reliably detect radio galaxy images from a telescope with different resolution and sensitivity. We find that the cold posterior effect persists in the models trained with iVON. Our results suggest that the choice of the optimiser can lead to qualitatively different solutions and future work using probabilistic neural network models should carefully consider the inductive biases being encoded through the optimisation process, in addition to the data, architecture and inference method.

1 Introduction

Upcoming radio astronomy surveys are expected to produce exabyte scale data which requires statistically robust machine learning models to extract scientific information [1]. Bayesian neural networks (BNNs) provide a principled way to model uncertainty in deep learning models. A recent benchmark of Bayesian deep learning for radio galaxy classification suggests that Hamiltonian Monte Carlo [HMC; 2, 3] performs well in terms of predictive accuracy, calibration of uncertainty and detecting distribution shift [4]. However, given the computational cost of HMC, the authors suggest that improving variational inference [VI; 5, 6] is a promising direction for future radio surveys. The authors use the Bayes by Backprop [BBB-VI; 7] algorithm to implement variational inference which performs reasonably well compared to all the other approximations considered but faces some challenges such as sensitivity to initialisation, slow convergence and the cold posterior effect (according to which the posterior must be down-weighted to get good predictive performance).

In variational inference, the negative Evidence Lower Bound (ELBO) is most commonly optimised using Stochastic Gradient Descent (SGD) based updates of the variational parameters. Gradient descent algorithms assume that the parameters can be treated as Cartesian coordinates in a Euclidean space, which is a reasonable assumption for standard point-wise neural networks. However, since VI

---

[*]The Alan Turing Institute, 96 Euston Rd, London, UK a.scaife@turing.ac.uk

is optimising a family of distributions over the neural network parameters, the parameter space forms a statistical manifold where every point on the manifold corresponds to a probability distribution over some domain $\mathcal{X}$. Distances on such manifolds are measured by defining a metric that induces an inner product on the tangent space at each point of the manifold. The space of all possible distributions of neural network parameters is a Riemannian manifold, with the Fisher information matrix (FIM) as its metric [8]. Therefore to account for the geometry of the variational parameters, it is possible to precondition the gradient with the inverse of the FIM. This technique is known as natural gradient descent (NGD). NGD can also be viewed through the lens of second order optimisation if the FIM is used to approximate the Hessian.

In this work we examine whether using NGD can improve variational inference based classification of radio galaxies. Although the FIM is intractable to calculate for large deep learning models, early work shows how adding adaptive weight noise to natural gradient for point estimation can be used to approximate natural gradient variational inference updates [Noisy Natural Gradient; 9]. Recent advances in the field of deep learning optimisation have led to the development of scalable variational learning algorithms like Improved Variational Online Newton based on natural gradient descent [iVON; 10, 11, 12]. We use the iVON optimiser and find that the posterior predictive distributions from iVON provide better calibrated uncertainties than all the approximate Bayesian inference methods considered in the previous benchmark , while providing similar predictive performance to the best performing models. However, this comes at the cost of reduced ability to detect distribution shift from near out-of-distribution (OoD) radio data, while preserving the ability to detect far-OoD data like images of optical galaxies.

2 Natural Gradient Descent for Variational Inference with iVON

Instead of optimising for point-wise model parameters, $\boldsymbol{\theta}$, with maximum likelihood or maximum a posteriori estimation (in case regularisation is used), a Bayesian neural network places a prior over the model parameters, $p(\boldsymbol{\theta})$, and infers a posterior probability distribution $p(\boldsymbol{\theta}|\mathcal{D})$ over the model parameters given data, $\mathcal{D}$, using Bayes rule. Since this posterior is intractable to compute for neural networks, various approximations are used [7, 13, 3, 14]. VI is one such approximation in which the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ is approximated with a parametric distribution, $q(\boldsymbol{\theta})$, from a family of distributions $\mathcal{Q}$ [15]. To find the optimal parameters of this variational distribution, the negative ELBO cost function is optimised:

$$\mathcal{L}(q) \equiv \min_{q(\boldsymbol{\theta}) \in \mathcal{Q}} -\mathbb{E}_{q(\boldsymbol{\theta})}[\log p(\mathcal{D}|\boldsymbol{\theta})] + \mathrm{KL}[q(\boldsymbol{\theta})||p(\boldsymbol{\theta})], \tag{1}$$

where $p(\mathcal{D}|\boldsymbol{\theta})$ is the likelihood. The improved Variational Online Newton (iVON) algorithm builds on the Bayesian Learning Rule [BLR; 11] for minimal exponential families. Many learning algorithms can be derived as a special case of the Bayesian Learning Rule (BLR) which optimises the following variational objective to approximate the posterior:

$$\mathcal{L}(q) \equiv \min_{q(\boldsymbol{\theta}) \in \mathcal{Q}} \mathbb{E}_{q(\boldsymbol{\theta})}[l(\boldsymbol{\theta})] - \mathcal{H}(q). \tag{2}$$

Here Equation 1 is re-written in terms of a generic loss function defined as $l(\boldsymbol{\theta}) = -\log p(\mathcal{D}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})$ and the entropy, $\mathcal{H}(q)$. If the variational distribution is chosen to be a minimal exponential family with natural parameters $\boldsymbol{\lambda}$ and expectation parameters $\boldsymbol{\mu}$, the above objective can be minimised by exploiting the information geometry of the posterior using natural gradients to update the natural parameters at current optimisation step, t, as follows [16]:

$$\boldsymbol{\lambda}_{t+1} \leftarrow \boldsymbol{\lambda}_t - \alpha\,\mathbf{F}^{-1}\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}), \tag{3}$$

where α is the learning rate and $\mathbf{F}$ is the FIM of the natural parameters. The bijection between the natural and expectation parameter space allows one to write the natural gradient with respect to the natural parameters as a gradient with respect to the expectation parameter and the following update is derived [11]:

$$\boldsymbol{\lambda}_{t+1} \leftarrow \boldsymbol{\lambda}_t - \alpha\nabla_{\boldsymbol{\mu}}\{\mathbb{E}_{q_{\boldsymbol{\lambda}_t}}[l(\boldsymbol{\theta})] - \mathcal{H}(q)\}. \tag{4}$$

In the following equations, we will denote $\mathbb{E}_{q_{\boldsymbol{\lambda}_t}}$ as $\mathbb{E}_{q_t}$ for clarity. By making different choices of the variational distribution $q(\boldsymbol{\theta})$ and different approximations to the natural gradient, one can derive gradient descent, Newton's method, Stochastic VI and many other learning algorithms [11].

When $q(\boldsymbol{\theta})$ is chosen to be a multivariate Gaussian, $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}, \mathbf{S}^{-1})$, BLR resembles Newton's update: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \mathbf{H}_{\boldsymbol{\theta}}[\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta})]$, where $\mathbf{H} = \nabla^2_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$ is the hessian of the loss. For a multivariate Gaussian, the natural parameters in terms of the mean $\mathbf{m}$ and precision $\mathbf{S}$ are $\boldsymbol{\lambda} := \{\mathbf{Sm}, -\mathbf{S}/2\}$ and the expectation parameters are $\boldsymbol{\mu} := \{\mathbb{E}_q(\boldsymbol{\theta}), \mathbb{E}_q(\boldsymbol{\theta}\boldsymbol{\theta}^T)\}$. Plugging the values of the first and second natural and expectation parameters in Equation 4 , the following update rules can be derived for the mean and the precision in terms of the gradient and Hessian of the loss with samples from $q(\boldsymbol{\theta})$ [11]:

$$\mathbf{m}_{t+1} \leftarrow \mathbf{m}_t - \alpha_t \mathbf{S}^{-1}_{t+1} \mathbb{E}_{q_t}[\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta})] \tag{5}$$

$$\mathbf{S}_{t+1} \leftarrow (1 - \alpha_t)\mathbf{S}_t + \alpha_t \mathbb{E}_{q_t}[\nabla^2_{\boldsymbol{\theta}} l(\boldsymbol{\theta})] \tag{6}$$

The precision update involves the hessian which is intractable to compute. The iVON algorithm makes further simplifications to scale the learning algorithm to large models and datasets. It makes a diagonal approximation to the hessian and uses the reparameterisation trick to calculate second order information using the mini-batch gradients $\hat{\mathbf{g}} = \hat{\nabla} l(\boldsymbol{\theta})$, $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}, \text{diag}(\boldsymbol{\sigma}))$:

$$\hat{\mathbf{h}} = \hat{\mathbf{g}} \, (\boldsymbol{\theta} - \mathbf{m})/\boldsymbol{\sigma}^2. \tag{7}$$

Thus iVON approximates second order curvature information by measuring how first order information (gradient) is affected by random perturbations of the parameters. An exponentially moving average of the gradient and hessian estimates is used to update the mean and standard deviation values. Assuming a multivariate Gaussian variational distribution implies that the precision matrix $\mathbf{S}$ must be positive-definite [12]. To handle the positive-definite constraint, an additional geometric correction term derived from approximate Riemannian gradient descent is added to the Hessian approximation, $\mathbf{h}$, to ensure its positivity:

$$\mathbf{h} \leftarrow (1 - \rho)\mathbf{h} + \rho\hat{\mathbf{h}} + \frac{1}{2}\rho^2(\mathbf{h} - \hat{\mathbf{h}})^2/(\mathbf{h} + s_0/\text{ess}), \tag{8}$$

where s_0 is the prior precision assuming $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|0, \mathbf{I}/s_0)$, ess is the effective sample size (ESS), and $\rho > 0$ is a constant. In practice, the weight decay regulariser, δ, is used instead of prior precision. Setting the ESS to the size of the training set, $\text{ess} = N$, is equivalent to setting the temperature of the posterior to 1. Setting $\text{ess} > N$ gives cold posteriors which may be required in practice to get good predictive performance [17, 18, 4]. The geometric correction term mathematically guarantees that the estimated precision remains positive and the variational distribution remains valid throughout training. The mean and standard deviation of the model parameters are then calculated as follows:

$$\mathbf{m}_{t+1} \leftarrow \mathbf{m}_t - \alpha \frac{(\hat{\mathbf{g}} + \delta\mathbf{m}_t)}{(\mathbf{h} + \delta)}; \qquad \boldsymbol{\sigma} \leftarrow \frac{1}{\sqrt{\text{ess}(\mathbf{h} + \delta)}}. \tag{9}$$

For the mean estimation, the numerator $(\hat{\mathbf{g}} + \delta\mathbf{m})$ gives the direction of update and the denominator $(\mathbf{h} + \delta)$ gives an adaptive estimate of the curvature for each parameter. The authors provide a PyTorch implementation of iVON which facilitates its use similar to standard first-order optimisation algorithms like Adam. The updates in Equation 9 can be contrasted with the BBB-VI updates which calculate separate gradients of the variational objective with respect to the variational parameters in the Euclidean space given learning rate α: $\mathbf{m} \leftarrow \mathbf{m} - \alpha\nabla_{\mathbf{m}}\mathcal{L}(\boldsymbol{\theta})$ and $\boldsymbol{\sigma} \leftarrow \boldsymbol{\sigma} - \alpha\nabla_{\boldsymbol{\sigma}}\mathcal{L}(\boldsymbol{\theta})$.

3 Experimental Setup

Data and architecture We use the MiraBest Confident dataset [19] which consists of Fanaroff-Riley type I (FR1) and type II (FRII) radio galaxies to train a LeNet like architecture as specified in the previous benchmark [4]. Additionally, we use the MIGHTEE [20] and GalaxyMNIST[2] datasets to evaluate the ability of the method to detect distribution shift. Details are recalled in the appendix for completeness.

iVON experiments We use the iVON optimiser [10] to train the model for classification. We set the initial learning rate to 0.2 and use a cosine annealing based learning rate scheduler to anneal the learning rate to 0 after a warm up period of 5 epochs. The optimisation process is sensitive to the initialisation of the hessian which is specified through the h_0 hyperparameter. We test different values of $h_0 = \{0.01, 0.1, 0.5, 1, 5\}$ and find the best results are achieved with $h_0 = 0.5$. The effective

---

[2]`https://github.com/mwalmsley/galaxy_mnist`

sample size (ESS) hyperparameter, ess, is used to set the temperature of the posterior. We experiment with different values of ESS, ess $= \{0.1N, N, 10N, 100N\}$ and find that using a cold posterior with ess $= 10N$ gives the best performance in terms of predictive performance and calibration of uncertainty. The weight decay is set to $\delta = 10^{-4}$, which is the same value used in the SGD based BBB-VI implementation of [4]. To train the model we use one Monte Carlo sample as recommended in the original implementation of iVON, which is also consistent with the BBB implementation of VI used in the previous benchmark. Results are reported for models trained for 1000 epochs using 10 different random seeds and random shuffling of data between training and validation samples. We use an early stopping criterion to select the model with the lowest validation loss during training. The epoch at which early stopping comes into effect is different for each of the 10 runs, ranging from epoch 189 to epoch 972. Training a single model takes 25 mins on an NVIDIA A100 GPU. Code for this work is available at `https://github.com/devinamhn/RadioGalaxies-BNNs`

In addition to experiments with iVON, we reproduce BBB-VI and HMC results from the previous benchmark for Bayesian deep learning for radio galaxy classification from [4] for comparison.

4   Results

We construct the posterior predictive distribution with 200 Monte Carlo samples and calculate the predictive accuracy, Uncertainty Calibration Error (UCE) and energy scores for detecting distribution shift. UCE is a measure of the *inherent* calibration of the Bayesian posterior predictive distribution, which we calculate using using the 64% credible intervals of the distributions. Evaluation metrics are recalled in Appendix B for completeness.

Experimental results for iVON are reported in Table 1 for two different ESS values, along with results reproduced from the previous benchmark for BBB-VI and HMC from [4]. We find that iVON with ess $= 10N$ has similar predictive performance to BBB-VI and better UCE than BBB-VI, but it converges much faster (25mins vs 40 mins). Setting ess $= 100N$ achieves similar predictive performance and uncertainty calibration to BBB-VI.

We plot the histograms of energy scores for the different test sets in Figure1 to examine the ability of iVON to detect distribution shift. Out-of-distribution (OoD) data is expected to have higher energy scores. We use the model with the lowest validation error from the experiments to calculate the energy scores. Energy plots for BBB-VI and HMC are reproduced from the previous benchmark [4] for comparison. We find that while iVON is able to clearly distinguish optical galaxy images, it cannot reliably detect distribution shifted radio galaxies from the MIGHTEE dataset compared to BBB-VI, see Figures 1b and 1c. HMC inference is still the most reliable in detecting distribution shift from different radio galaxy datasets, see Figure 1a. One might expect the Mirabest test set to have the lowest energy score since the model was trained on radio galaxies from the same dataset, however, MIGHTEE has even lower energy scores. This is because MIGHTEE contains images of radio galaxies from a telescope with different resolution and sensitivity but the types of objects in the dataset are similar to the training data (i.e, FRI and FRII galaxies). So these objects are *near-OoD*, rather than *far-OoD* like the optical galaxies from GalaxyMNIST which have completely different morphological features and have a more significant distribution shift. Near-OoD detection is a more challenging problem than detecting far OoD data.

Table 1: Test error, uncertainty calibration error (UCE) of the predictive entropy and time to train for HMC, BBB-VI and iVON based Bayesian neural networks. The HMC and BBB-VI results are reproduced from [4]. For iVON, results are shown for two different values of the effective sample size (ess) as a multiple of the length of training data, N.

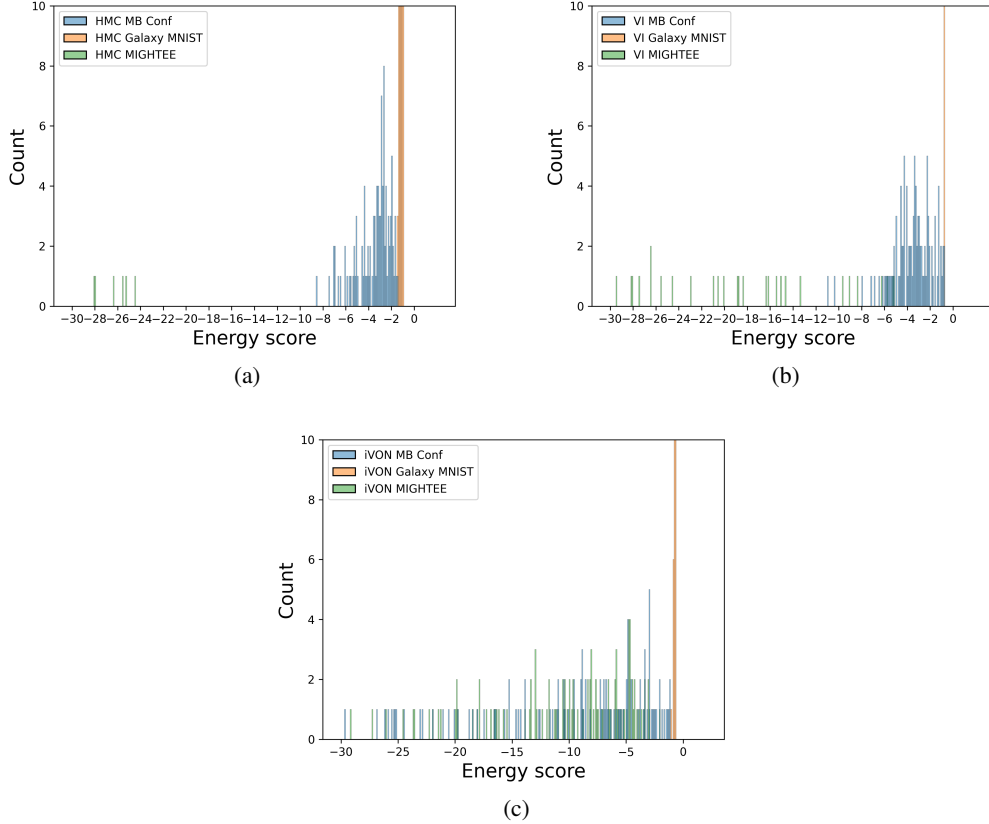| Inference	Error (%) ↓	UCE ↓	Time
HMC	4.16 ± 0.45	14.76 ± 0.95	7 days
BBB-VI	3.94 ± 0.01	12.77 ± 6.11	40 mins
iVON ess $= 10N$	3.07 ± 1.47	8.37 ± 4.12	25 mins
iVON ess $= 100N$	3.36 ± 1.23	12.19 ± 6.57	25 mins

Figure 1: Histograms of energy scores calculated for the MiraBest (MBConf; blue), GalaxyMNIST (orange) and MIGHTEE (green) test datasets for (a) HMC (b) BBB-VI and (c) iVON. The HMC and BBB-VI plots are reproduced from [4].

5   Conclusions and Discussion

We find that accounting for the geometry of the variational parameters through second order informa-tion provides a better conditioned objective for optimising the probability distributions being inferred by VI. For overparameterised models, different optimisers can converge to different optimal solutions while minimising the same objective function. Our results suggest that optimising the variational objective with iVON, which is a natural gradient descent based optimiser, provides better uncertainty calibration than SGD optimisation, and improved convergence speed while providing the same predictive performance. However, the ability to distinguish between different radio galaxy datasets is reduced compared to BBB-VI. This suggests that even if an optimiser converges to a reasonable solution faster, as is the case when we use NGD instead of SGD to optimise VI, it might not find a solution that will be good for a diverse range of objectives. Different optimisers induce different inductive biases [21]. They constrain the space of solutions that is reachable by any optimiser within the space of all solutions defined by the data, architecture, initialisation, which can lead to different types of representations (distributed and redundant vs localised and compressed, for example). Future work could look at formalising these trade-offs to induce the most useful inductive biases for a particular application domain.

Acknowledgments and Disclosure of Funding

References

[1] Tao An. Science opportunities and challenges associated with ska big data. *Science China Physics, Mechanics & Astronomy*, 62:1–6, 2019.

[2] Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

[3] Adam D Cobb and Brian Jalaian. Scaling hamiltonian monte carlo inference for bayesian neural networks with symmetric splitting. In *Uncertainty in Artificial Intelligence*, pages 675–685. PMLR, 2021.

[4] Devina Mohan and Anna M. M. Scaife. Evaluating bayesian deep learning for radio galaxy classification. In Negar Kiyavash and Joris M. Mooij, editors, *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 2587–2597. PMLR, 15–19 Jul 2024.

[5] Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28:2575–2583, 2015.

[6] Alex Graves. Practical variational inference for neural networks. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, page 2348–2356, Red Hook, NY, USA, 2011. Curran Associates Inc.

[7] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Networks. *arXiv e-prints*, page arXiv:1505.05424, May 2015.

[8] Shun-ichi Amari, Ryo Karakida, and Masafumi Oizumi. Fisher information and natural gradient learning in random deep networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 694–702. PMLR, 2019.

[9] Guodong Zhang, Shengyang Sun, David Duvenaud, and Roger Grosse. Noisy natural gradient as variational inference. In *International conference on machine learning*, pages 5852–5861. PMLR, 2018.

[10] Yuesong Shen, Nico Daheim, Bai Cong, et al. Variational learning is effective for large deep networks. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[11] Mohammad Emtiyaz Khan and Håyvard Rue. The bayesian learning rule. *Journal of Machine Learning Research*, 24(281):1–46, 2023.

[12] Wu Lin, Mark Schmidt, and Mohammad Emtiyaz Khan. Handling the positive-definite constraint in the Bayesian learning rule. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6116–6126. PMLR, 13–18 Jul 2020.

[13] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015.

[14] Erik Daxberger, Agustinus Kristiadi, Alexander Immer, et al. Laplace redux-effortless bayesian deep learning. *Advances in neural information processing systems*, 34:20089–20103, 2021.

[15] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *arXiv e-prints*, page arXiv:1601.00670, January 2016.

[16] Mohammad Emtiyaz Khan. Information geometry of variational bayes: Me khan. *Information Geometry*, pages 1–15, 2025.

[17] Florian Wenzel, Kevin Roth, Bastiaan S Veeling, et al. How good is the bayes posterior in deep neural networks really? *arXiv preprint arXiv:2002.02405*, 2020.

[18] Devina Mohan, Anna MM Scaife, Fiona Porter, et al. Quantifying uncertainty in deep learning approaches to radio galaxy classification. *Monthly Notices of the Royal Astronomical Society*, 511(3):3722–3740, 2022.

[19] Fiona AM Porter and Anna MM Scaife. Mirabest: a data set of morphologically classified radio galaxies for machine learning. *RAS Techniques and Instruments*, 2(1):293–306, 2023.

[20] I Heywood, MJ Jarvis, CL Hale, et al. Mightee: total intensity radio continuum imaging and the cosmos/xmm-lss early science fields. *Monthly Notices of the Royal Astronomical Society*, 509(2):2150–2168, 2022.

[21] Razvan Pascanu, Clare Lyle, Ionut-Vlad Modoranu, et al. Optimizers qualitatively alter solutions and we should leverage this. *arXiv preprint arXiv:2507.12224*, 2025.

[22] B. L. Fanaroff and J. M. Riley. The morphology of extragalactic radio sources of high and low luminosity. *mnras*, 167:31P–36P, May 1974.

[23] L. Rudnick and F. N. Owen. Head-tail radio sources in clusters of galaxies. *apjl*, 203:L107–L111, February 1976.

[24] C. P. O'Dea and F. N. Owen. VLA observations of 57 sources in clusters of galaxies. *aj*, 90:927–953, June 1985.

[25] Gopal-Krishna and P. J. Wiita. Extragalactic radio sources with hybrid morphology: implications for the Fanaroff-Riley dichotomy. *aap*, 363:507–516, November 2000.

[26] Arno P. Schoenmakers, A. G. de Bruyn, H. J. A. Röttgering, et al. Radio galaxies with a 'double-double morphology' - I. Analysis of the radio properties and evidence for interrupted activity in active galactic nuclei. *mnras*, 315(2):371–380, June 2000.

[27] Bernard L Fanaroff and Julia M Riley. The morphology of extragalactic radio sources of high and low luminosity. *Monthly Notices of the Royal Astronomical Society*, 167(1):31P–36P, 1974.

[28] Inigo V Slijepcevic, Anna MM Scaife, Mike Walmsley, et al. Radio galaxy zoo: towards building the first multipurpose foundation model for radio astronomy with self-supervised learning. *RAS Techniques and Instruments*, 3(1):19–32, 2024.

[29] Mike Walmsley, Chris Lintott, Tobias Géron, et al. Galaxy zoo decals: Detailed visual morphology measurements from volunteers and deep learning for 314 000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 509(3):3966–3988, 2022.

[30] Yarin Gal. Uncertainty in deep learning. PhD thesis, University of Cambridge, 2016.

[31] Max-Heinrich Laves, Sontje Ihler, Karl-Philipp Kortmann, and Tobias Ortmaier. Well-calibrated model uncertainty with temperature scaling for dropout variational inference. *arXiv preprint arXiv:1909.13550*, 2019.

[32] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

A Data

Radio galaxies are characterised by large scale jets and lobes which can extend up to mega-parsec distances from the central black hole and are observed in the radio spectrum. The original binary classification scheme proposed to classify such extended radio sources was based on the ratio of the extent of the highest surface brightness regions to the total extent of the galaxy [22]. Fanaroff Riley Type I (FRI) galaxies are edge-darkened whereas Fanaroff Riley Type II (FRII) galaxies are edge-brightened. Over the years, several other morphologies such as bent-tail [23, 24], hybrid [25], and double-double [26] sources have also been observed and there is still a continuing debate about the exact interplay between extrinsic effects, such as the interaction between the jet and the environment, and intrinsic effects, such as differences in central engines and accretion modes, that give rise to the different morphologies.

We train our BNNs on the MiraBest Confident dataset [Section A.1] and use the MIGHTEE [Section A.2] and GalaxyMNIST [Section A.3] datasets to test the ability of our BNNs to detect different types of distribution shifts.

A.1 MiraBest Dataset

The MiraBest dataset used in this work consists of 1256 images of radio galaxies of 150×150 pixels pre-processed to be used specifically for deep learning tasks [19]. The galaxies are labelled using the FRI and FRII morphological types based on the definition of [27] and further divided into their subtypes. In addition to labelling the sources as FRI, FRII and their subtypes, each source is also flagged as 'Confident' or 'Uncertain' to indicate the human classifiers' confidence while labelling the dataset. In this work we use the MiraBest Confident subset and consider only the binary FRI/FRII classification. The training and validation sets are created by splitting the predefined training data into a ratio of 80:20. The final split consists of 584 training samples, 145 validation samples, and 104 withheld test samples. Data augmentation in the form of random rotations is used.

A.2 MIGHTEE

The MIGHTEE dataset is constructed using the Early Science data products from the MeerKAT International GHz Tiered Extragalactic Exploration survey [MIGHTEE; 20]. MIGHTEE is an ongoing radio continuum survey being conducted using the MeerKAT telescope, which is one of the precursors to the Square Kilometer Array (SKA). The survey provides radio continuum, spectral line and polarisation data, of which we use the radio continuum data and extract images for the COSMOS and XMMLSS fields. While there are thousands of objects in these fields, expert labels are only available for 117 objects. We use the data pre-processing and expert labels made available by [28]. The dataset contains classifications based on the consensus of five expert radio astronomers. The final sample contains 45 FRI and 72 FRII galaxies We note that the MIGHTEE dataset contains significant observational differences from the MiraBest dataset.

A.3 Galaxy MNIST

In addition to considering different datasets of radio galaxies which have been curated using using data from radio telescopes, we also evaluate our models on data collected from optical telescopes. Optical images of galaxies contain different features and in a sense represent completely out-of-distribution galaxies which well-calibrated models should classify with a very high degree of uncertainty so that they can be flagged for inspection by an expert. We use the GalaxyMNIST[3] dataset which contain images of $10,000$ optical galaxies classified into four morphological types using labels collected by the Galaxy Zoo citizen science project The galaxies are drawn from the Galaxy Zoo Decals catalogue [29]. We resize the high resolution images from 224x224 to 150x150 to match the input dimensions of our model. We construct a small test set of 104 galaxies from the dataset to evaluate the out-of-distribution detection ability of our BNNs.

---

[3]`https://github.com/mwalmsley/galaxy_mnist`

B   Evaluation metrics

We use the expected value of the posterior predictive distribution to obtain the classification of each galaxy in the MiraBest test set and calculate the test error for a single experimental run by taking an average of the classification error over the entire test set. We report the mean and standard deviation of the test error for 10 experimental runs.

B.1   Predictive entropy

Using Monte Carlo (MC) samples obtained from the posterior predictive distributions of different BNNs, one can obtain N Softmax probabilities for each class, c, in the dataset. We can recover N class-wise Softmax probabilities as follows:

$$P(y|x, D) = \frac{1}{N} \sum_{i=1}^{N} P(y = c|x, w^{(i)}), \tag{10}$$

where $w^{(i)}$ is the ith weight sample from the posterior distribution conditioned on the training data, D, and (x, y) are data from the test set. Using these samples, one can quantify the uncertainty in the predictions using different metrics. In this work we look at the predictive entropy of the softmax distribution which measures the average amount of information inherent in the distribution and is defined as:

$$\mathbb{H}(y|x, D) = -\sum_{c} P(y = c|x, w) \log P(y = c|x, w), \tag{11}$$

which can be approximated using MC samples as follows [30]:

$$\mathbb{H}(y|x, D) = -\sum_{c} \left(\frac{1}{N} \sum_{i=1}^{N} P(y = c|x, w^{(i)}) \right) \log \left(\frac{1}{N} \sum_{i=1}^{N} P(y = c|x, w^{(i)}) \right). \tag{12}$$

B.2   Uncertainty Calibration Error (UCE)

We report the expected uncertainty calibration error [UCE; 13, 31, 18] of the predictive entropy of our posterior predictive distributions in Table 1. We use the 64% credible intervals of the posterior predictive distributions to calculate UCE. To examine how well calibrated the predictive entropy is, we calculate the Uncertainty Calibration Error (UCE), which is a more general form of the Expected Calibration Error (ECE). UCE is a weighted average of the difference between fractional error and uncertainty calculated for the output of the model when binned into M bins of equal width for a particular uncertainty metric:

$$\text{UCE} = \sum_{m=1}^{M} \frac{|B(m)|}{n} |\text{err}(B_m) - \text{uncert}(B_m)|. \tag{13}$$

Here B_m is the set of data in a particular bin, n is the total number of data points and $\text{uncert}(B_m)$ is the average value of a given uncertainty metric for those data points:

$$\text{uncert}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \text{uncert}_i , \tag{14}$$

where uncert_i can be calculated using Equation 12 followed by minmax-normalisation to bring values into the range 0 to 1.

Equation 16 of [31] defines the average fractional error in bin B_m to be:

$$\text{err}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \text{err}_i , \tag{15}$$

where err_i is the contribution to this error from an individual data point, defined as:

$$\text{err}_i = \mathbf{1}(\hat{y}_i \neq y) \quad \forall \ i \in B_m . \tag{16}$$

Here we redefine err_i to be the average error obtained for an individual data sample, such that

$$\text{err}_i = \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}(\hat{y}_{ij} \neq y) \quad \forall \ i \in B_m , \tag{17}$$

where N is the number of samples drawn from the posterior predictive distribution.

B.3 Energy score

[32] propose a post-hoc scoring function for discriminative classification models which can be used to distinguish between in-distribution (iD) and distribution-shifted data examples. We calculate average scalar energy scores for different test samples, x, for all the datasets considered in this work using the logit values, $f_i(x)$, for each class, i, using N posterior samples:

$$\tilde{\mathrm{E}}(x; f) = \frac{1}{N} \sum_j^N -T. \log \sum_i^K e^{f_i(x)/T}, \tag{18}$$

where the temperature term, T, is set to 1. Out-of-distribution (OoD) samples are expected to have higher energy in this framework.