MCMC to address model misspecification in Deep Learning 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

The radio astronomy community is adopting deep learning techniques to deal with the huge data volumes expected from the next-generation of radio observatories. Bayesian neural networks (BNNs) provide a principled way to model uncertainty in the predictions made by deep learning models and will play an important role in extracting well-calibrated uncertainty estimates from the outputs of these models. However, most commonly used approximate Bayesian inference techniques such as variational inference and MCMC-based algorithms experience a "cold posterior effect (CPE)", according to which the posterior must be down-weighted in order to get good predictive performance. The CPE has been linked to several factors such as data augmentation or dataset curation leading to a misspecified likelihood and prior misspecification. In this work we use MCMC sampling to show that a Gaussian parametric family is a poor variational approximation to the true posterior and gives rise to the CPE previously observed in morphological classification of radio galaxies using variational inference based BNNs.

1 Introduction

The next-generation of radio astronomy facilities such as the Square Kilometre Array (SKA) will produce huge volumes of data and the use of deep learning (DL) methods is inevitable given the expected data volumes [1, 2]. Modern astrophysics is driven by population analyses and any automated classification pipeline should produce well-calibrated uncertainty estimates that quantify the model uncertainty introduced in the results. In this work we consider the morphological classification of radio galaxies and discuss the challenges faced while implementing Bayesian Convolutional Neural Networks (CNNs) for their classification.

While several works have looked at classifying radio galaxies with deep learning [e.g. 3, 4, 5, 6, 7], with the exception of [8] and [9], little work has been done on understanding the degree of confidence with which CNN models predict the class of individual radio galaxies. In general, it has been suggested that deep learning models produce overconfident predictions [10] and provide no uncertainty estimates, which are essential for scientific application of these models. On the other hand, probabilistic models such as Bayesian neural networks (BNNs) provide a principled way to model uncertainty [11, 12] by specifying priors, $P(\theta)$, over the neural network parameters, θ, and learning the posterior distribution, $P(\theta|D)$, over those parameters, where D is the data.

Recovering this posterior distribution directly is intractable for neural networks. Several techniques have been developed to approximate Bayesian inference for neural networks among which Variational Inference (VI) and Monte Carlo (MC) Dropout methods are most commonly used. VI assumes an approximate posterior from a family of tractable distributions, and converts the inference problem

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

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

into an optimisation problem [13, 14, 15]. The model learns the parameters of the distributions by minimising an Evidence Lower Bound Objective (ELBO) function.

Another easily implemented Bayesian approximation is MC Dropout, which learns a distribution over the network outputs by setting randomly selected weights of the network to zero with probability, p [16]. MC dropout can be considered an approximation to VI, where the variational approximation is a Bernoulli distribution. Although a convenient technique, this method lacks flexibility and does not fully capture the uncertainty in model predictions, especially under covariate shift where the data distributions at training and test time are not identically distributed [17].

However, there are several challenges in implementing BNNs in practice. Several published works have reported that their BNNs experience a "cold posterior effect (CPE)", according to which the posterior needs to be down-weighted or tempered with a temperature term, $T \leq 1$, in order to get good predictive performance [18]:

$$P(\theta|D) \propto (P(D|\theta)P(\theta))^{1/T}. \tag{1}$$

Previous work has shown that VI based BNN models experience a CPE when classifying radio galaxies [9]. The choice of variational approximation limits the variational posterior to specific regions of the true posterior density space and it is difficult to evaluate how good the variational approximation is without having access to the true posterior [19]. Several hypothesis have been put forward to explain the CPE including likelihood, prior and model misspecification [18, 20, 21].

In this work we demonstrate that, for radio galaxy classification, using MCMC to recover posterior distributions on neural network parameters suggests that the "cold posterior effect" previously observed with VI models is due to model misspecification arising from poor variational posterior approximations. We also compare model performance for different approximate Bayesian inference methods including VI and MC Dropout and present preliminary uncertainty calibration results.

2 MCMC for Neural Networks

MCMC methods are a class of algorithms used to obtain samples from probability distributions which are otherwise intractable or do not have a full analytical description. The first application of MCMC to neural networks was proposed by [22], who introduced Hamiltonian Monte Carlo (HMC) from quantum chromodynamics to the general statistics literature. However, it wasn't until [23] introduced Stochastic Gradient Langevin Dynamics (SGLD), that MCMC for neural networks became feasible for large datasets. More recently, [24] have revisited HMC and proposed novel data splitting techniques to make it work with large datasets. We use the HMC algorithm in our work.

Hamiltonian Monte Carlo HMC simulates the path of a particle traversing the negative posterior density space using Hamiltonian dynamics [25, 26, 27]. To apply HMC to deep learning, the neural network parameter space is augmented by specifying an additional momentum variable, m, for each parameter, θ. Therefore, for a d-dimensional parameter space, the augmented parameter space contains $2d$ dimensions. We can then define a log joint density as follows:

$$\log[p(\theta, m)] = \log[p(\theta|D)p(m)]\,. \tag{2}$$

Hamiltonian dynamics allows us to travel on the contours defined by the joint density of the position and momentum variables, also known as the phase space. The Hamiltonian function is given by:

$$H(\theta, m) = U(\theta) + K(m) = constant, \tag{3}$$

where $U(\theta)$ is the potential energy and $K(m)$ is the kinetic energy. The potential energy is defined to be the negative log posterior probability and the kinetic energy is usually assumed to be quadratic in nature and of the form $K(m) = (1/2)\, m^T M^{-1} m$, where M is a positive-definite mass matrix. This corresponds to the negative probability density of a zero-mean Gaussian, $p(m) = \mathcal{N}(m|0, M)$, with covariance matrix, M, which is usually assumed to be the identity matrix.

The partial derivatives of the Hamiltonian describe how the system evolves with time. In order to solve the partial differential equations using computers, we need to discretise the time, t, of the dynamical simulation using a step-size, ϵ. The state of the system can then be computed iteratively at times ϵ, 2ϵ, 3ϵ... and so on, starting at time zero upto a specified number of steps, L. The leapfrog

integrator is used to solve the system of partial differential equations. Two hyperparameters, the step-size, ϵ, and the number of leapfrog steps, L, together determine the trajectory length of the simulation. The partial derivative of the potential energy with respect to the position, $\partial U/\partial\theta$, can be calculated using the automatic differentiation capabilities of most standard neural network libraries.

In each iteration of the HMC algorithm, new momentum values are sampled from Gaussian distributions, followed by simulating the trajectory of the particles according to Hamiltonian dynamics for L steps using the leapfrog integrator with step-size ϵ. At the end of the trajectory, the final position and momentum variables, (θ^*, m^*), are accepted based on a Metropolis-Hastings accept/reject criterion that evaluates the Hamiltonian for the proposed parameters and the previous parameters.

3 Experimental Setup

Data 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 [28]. The galaxies are labelled using the FRI and FRII morphological types based on the definition of [29] 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. No data augmentation is used.

Architecture We use an expanded LeNet-5 architecture with two additional convolutional layers with 26 and 32 channels, respectively, to be consistent with the literature on using BNNs for classifying the MiraBest dataset [9]. The model has $232,444$ parameters in total.

MCMC Inference We use the HAMILTORCH package[2] developed by [24] for scaling HMC to large datasets. Using their HMC sampler, we set up two HMC chains of $200,000$ steps using different random seeds and run it on the MiraBest Confident dataset. We use a step size of $\epsilon = 10^{-4}$ and set the number of leapfrog steps to $L = 50$. We specify a Gaussian prior over the network parameters and evaluate different prior widths, $\sigma = \{1, 10^{-1}, 10^{-2}, 10^{-3}\}$, using the validation data set. We find that $\sigma = 10^{-1}$ results in the best predictive performance and consequently use it to define the prior width for all weights and biases of the neural network in our experiments. A burn-in of $20,000$ samples is discarded. To compute the final posteriors we thin the chains by a factor of 1000 to reduce the autocorrelation in the samples and obtain 180 samples. A compute time of 150 hrs is required to run the inference on two Nvidia A100 GPUs. The Gelman-Rubin diagnostic, $\hat{R}$, is used to assess the convergence of our HMC chains [30]. If $\hat{R} \approx 1$ we consider the MCMC chains for that particular parameter to have converged. While $\hat{R}$ values for some parameters in the network are greater than 1, the final two neurons in the last layer of our network have $\hat{R} \leq 1$. We also monitor the negative log-likelihood and accuracy, which converge by the $100,000^{\text{th}}$ inference step.

Other models For the VI implementation we use a Gaussian variational approximation to the posterior and consider different priors including Gaussian and Laplace distributions following [9]. The Laplace prior provides optimal predictive performance and lowest uncertainty calibration error, however for direct comparison to our HMC baseline we also consider a Gaussian prior with $\sigma = 0.01$. Results are reported for a tempered VI posterior, with $T = 0.01$ in Table 1. For the MC Dropout model, a dropout rate of 50% is implemented before the last layer of our neural network, which is standard for CNNs [8, 16]. The network is trained for 150 epochs using the Adam optimser with a learning rate of 10^{-3} and a weight decay of 10^{-4}. We obtain 200 samples from VI and MC Dropout posterior predictive distributions by passing each sample in the test set through the test loop 200 times. We use the same optimiser hyperparameters as the MC Dropout training for our non-Bayesian CNN model. A compute time of 12 mins is required to train the VI model on a single Nvidia A100 GPU.

Code for this work is available at `https://github.com/devinamhn/RadioGalaxies-BNNs`

[2] `https://github.com/AdamCobb/hamiltorch`

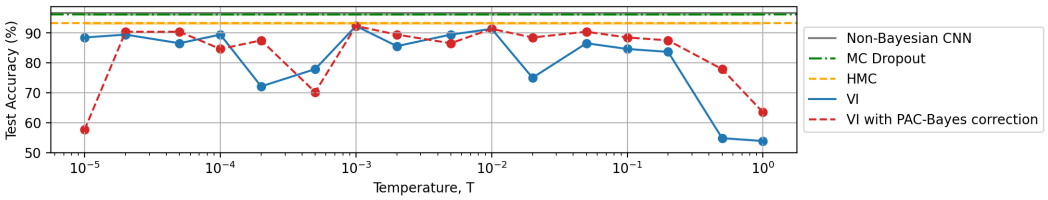

Figure 1: The "cold posterior" effect (CPE) observed in VI models (solid blue line) persists inspite of a PAC-Bayes correction term to account for model misspecification (red dashed line) when classifying radio galaxies. Using samples from HMC, we demonstrate that using a Gaussian variational distribution leads to a poor approximation to the posterior, giving rise to the CPE. Data are also shown for MC Dropout (green dot-dashed line) and our non-Bayesian CNN (solid gray line) for comparison.

Table 1: Test error and class-wise expected uncertainty calibration error (cUCE) for predictive entropy are reported for different BNNs for the MiraBest Confident dataset.

Model	**Error (%)**	**% cUCE**
Non-Bayesian CNN	3.33	$\times$
MC Dropout	3.85 ± 0.19	9.75
VI (Laplace prior)	10.58 ± 0.30	14.26
VI (Gaussian prior)	12.96 ± 0.33	30.05
HMC	6.73 ± 0.25	13.17

4 Results

Cold posterior effect Previous work on using VI for radio galaxy classification has shown that the "cold posterior effect" (CPE) persists even when the learning strategy is modified to compensate for model misspecification with a second order PAC-Bayes bound to improve the generalisation performance of the network [9, 31], see Figure 1. We do not observe a CPE when we use samples from our HMC inference to construct the posterior predictive distribution for classifying the MiraBest dataset (orange dashed line in Figure 1). This suggests that using a Gaussian parametric family as a variational approximation to the true posterior distribution is a poor assumption and leads to a misspecified model, which gives rise to the CPE. In the general Bayesian DL literature, some authors argue that CPE is mainly an artifact of data augmentation [20], while others have shown that data augmentation is a sufficient but not necessary condition for CPE to be present [21]. We find that data augmentation does not have a significant effect on our HMC and VI models.

Model performance The test error is calculated by taking an average of the predictions obtained using the expected value of the posterior predictive distribution for each galaxy in the MiraBest Confident test set for different models, see Table 1. No data augmentation is used during training/inference. The non-Bayesian CNN and MC Dropout perform comparably in terms of test error. HMC is more accurate than VI, but does not match the predictive performance of MC Dropout. We note that it is not performance alone that is important for our application, but also the calibration of the posterior uncertainties which will influence the scientific analysis performed using the catalogues generated by DL pipelines. We have conducted preliminary uncertainty calibration experiments using the 64% credible intervals of the posterior predictive distributions to calculate the class-wise expected Uncertainty Calibration Error (cUCE) values for the predictive entropy [9, 16, 32]. However, at this stage we do not draw any strong conclusions from the uncertainty quantification experiments, which will be considered more fully in our future work.

5 Conclusions

Using samples from HMC, we find that the cold posterior effect previously observed in the morphological classification of radio galaxies using variational inference arises from using a misspecified parametric family to approximate the posterior. While MCMC does not provide the most computationally efficient framework for approximate Bayesian inference for neural networks, it produces

asymptotically exact samples from the posterior which are useful for developing more accurate approximate Bayesian inference techniques for the radio galaxy classification problem in future.

Acknowledgments and Disclosure of Funding

AMS gratefully acknowledges support from an Alan Turing Institute AI Fellowship EP/V030302/1.

References

[1] AMM Scaife. Big telescope, big data: towards exascale with the square kilometre array. *Philosophical Transactions of the Royal Society A*, 378(2166):20190060, 2020.

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

[3] A K Aniyan and K Thorat. Classifying Radio Galaxies with the Convolutional Neural Network. *The Astrophysical Journal Supplement Series*, 230(2):20, 2017.

[4] V Lukic, M Brüggen, Beatriz Mingo, et al. Morphological classification of radio galaxies: capsule networks versus convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, 487(2):1729–1744, 2019.

[5] H Tang, A M M Scaife, and J P Leahy. Transfer learning for radio galaxy classification. *Monthly Notices of the Royal Astronomical Society*, 488(3):3358–3375, 07 2019.

[6] Micah Bowles, Anna MM Scaife, Fiona Porter, et al. Attention-gating for improved radio galaxy classification. *Monthly Notices of the Royal Astronomical Society*, 501(3):4579–4595, 2021.

[7] Inigo V Slijepcevic, Anna MM Scaife, Mike Walmsley, et al. Radio galaxy zoo: using semi-supervised learning to leverage large unlabelled data sets for radio galaxy classification under data set shift. *Monthly Notices of the Royal Astronomical Society*, 514(2):2599–2613, 2022.

[8] Anna MM Scaife and Fiona Porter. Fanaroff–riley classification of radio galaxies using group-equivariant convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, 503(2):2369–2379, 2021.

[9] 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.

[10] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[11] David J. C. MacKay. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 4(3):448–472, 05 1992.

[12] David J. C. MacKay. The evidence framework applied to classification networks. *Neural Computation*, 4(5):720–736, 1992.

[13] 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.

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

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

[17] Alex Chan, Ahmed Alaa, Zhaozhi Qian, and Mihaela Van Der Schaar. Unlabelled data improves bayesian uncertainty calibration under covariate shift. In *International Conference on Machine Learning*, pages 1392–1402. PMLR, 2020.

[18] 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.

[19] Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Yes, but did it work?: Evaluating variational inference. In *International Conference on Machine Learning*, pages 5581–5590. PMLR, 2018.

[20] Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. What are bayesian neural network posteriors really like? In *International conference on machine learning*, pages 4629–4640. PMLR, 2021.

[21] Lorenzo Noci, Kevin Roth, Gregor Bachmann, et al. Disentangling the roles of curation, data-augmentation and the prior in the cold posterior effect. *Advances in Neural Information Processing Systems*, 34:12738–12748, 2021.

[22] Radford M Neal and Geoffrey E Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

[23] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

[24] 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.

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

[26] Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

[27] David W Hogg and Daniel Foreman-Mackey. Data analysis recipes: Using markov chain monte carlo. *The Astrophysical Journal Supplement Series*, 236(1):11, 2018.

[28] 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.

[29] 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.

[30] Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992.

[31] Andrés R Masegosa. Learning under model misspecification: Applications to variational and ensemble methods. *arXiv preprint arXiv:1912.08335*, 2019.

[32] 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.