Calibrating Bayesian Neural Networks with Alpha-divergences and Normalizing Flows

Héctor J. Hortúa, Luigi Malagò, Riccardo Volpi

Romanian Institute of Science and Technology, Cluj Napoca, Romania
email: volpi@rist.ro

UNIUNEA EUROPEANĂ

Instrumente Structurale 2014-2020

Abstract

Bayesian Neural Networks are successfully employed for confidence intervals estimation and they are able to provide two types of uncertainties: aleatoric, due to the intrinsic noise of the data, and epistemic, due to the ignorance of the model. Nevertheless BNNs are usually uncalibrated after training and tend towards overconfidence in predicting the errors. An effective method for calibrating BNNs should effectively calibrate both epistemic and aleatoric uncertainties, with low impact in terms of computational complexity. In this work we show how BNN calibration is related to the value of the negative log likelihood and employing several approaches as normalizing flows and alpha divergence we can obtain well-calibrated BNNs. We empirically demonstrate the advantages of these techniques in regression problems involving parameters estimation with correlations between their output uncertainties. We compare the use of normalizing flows and alpha divergence both in training and in the calibration phase, and we show how the latter provides more reliable uncertainty estimates for specific choices of alpha, a better coefficient of determination, and it is a considerably more efficient approach especially for complex network architectures. Furthermore we apply the presented framework to adversarial examples for traffic signs recognition and show its robustness advantage.

1 Methods

The variational inference approach consists in approximating the posterior distribution $p(\mathbf{w}|\mathcal{D})$ with a variational distribution $q(\mathbf{w}|\theta)$, depending on a set of parameters θ [2, 6]. The objective can then be formalized as finding θ that makes q as close as possible to the true posterior, for instance by minimizing the KullBack-Leibler (KL) divergence between the two distributions [6]

$$\mathrm{KL}(q(\mathbf{w}|\theta)||p(\mathbf{w}|\mathcal{D})) \equiv \int_\Omega q(\mathbf{w}|\theta) \ln \frac{q(\mathbf{w}|\theta)}{p(\mathbf{w}|\mathcal{D})} d\mathbf{w}. \tag{1}$$

Using Bayes theorem, we can find that minimizing Eq. (1) is equivalent to minimizing

$$\mathrm{KL}(q(\mathbf{w}|\theta)||p(\mathbf{w})) - \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \int_\Omega q(\mathbf{w}|\theta) \ln p(\mathbf{y}|\mathbf{x},\mathbf{w}) d\mathbf{w}. \tag{2}$$

If the network is minimized at $\hat{\theta}$, the probability distribution of y^* for a new input x^* can be written as [6]

$$q_{\hat{\theta}}(\mathbf{y}^*|\mathbf{x}^*) = \int_\Omega p(\mathbf{y}^*|\mathbf{x}^*,\mathbf{w}) q(\mathbf{w}|\hat{\theta}) d\mathbf{w}, \tag{3}$$

while the covariance of the variational predictive distribution, for a fixed $\boldsymbol{x}^*$ is [6, 5]

$$\mathrm{Cov}_{q_{\hat{\theta}}}(\mathbf{y}^*,\mathbf{y}^*|\mathbf{x}^*) \equiv \mathbb{E}_{q_{\hat{\theta}}}[\mathbf{y}^*\mathbf{y}^{*\mathrm{T}}|\mathbf{x}^*] - \mathbb{E}_{q_{\hat{\theta}}}[\mathbf{y}^*|\mathbf{x}^*]\mathbb{E}_{q_{\hat{\theta}}}[\mathbf{y}^*|\mathbf{x}^*]^{\mathrm{T}}. \tag{4}$$

Alternatively we can consider minimizing the α divergence [1, 3] defined as

$$\mathrm{D}_\alpha[p||q] = \frac{1}{\alpha(1-\alpha)} \left(1 - \int p(x)^\alpha q(x)^{1-\alpha} dx\right), \tag{5}$$

where $\alpha = 0$ is the KL used in VI, $\alpha = 1.0$ is used in EP, the case $\alpha = 0.5$ is the Hellinger distance and $\alpha = 2$ is the χ^2 distance. In the limit of $\alpha/D \to 0$, the authors in [3, 7] arrive to a generalization of Eq. 2 given by

$$\begin{aligned}\mathcal{L}_\alpha \approx\ & \mathrm{KL}(q(\boldsymbol{w}|\boldsymbol{\theta})||p(\boldsymbol{w})) \\ & - \frac{1}{\alpha} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \ln \int_\Omega q(\mathbf{w}|\boldsymbol{\theta}) p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})^\alpha d\boldsymbol{w}\ ,\end{aligned} \tag{6}$$

allowing to optimize the family of α divergences, determining approximate distributions q with different properties. We use a variant of VGG16 architecture, we train with Adam-optimizer. averaged over the mini-batch. To model the distribution over the weights we used two methods, Dropout [6] and Flipout [11]. For Dropout we tested several dropout rates while keeping L2 regularization fixed to $1e^{-5}$, while for Flipout we tested several L2 regularizations.

2 The 21cm signal

The 21cm signal from the neutral hydrogen in the intergalactic medium (IGM) is described through its brightness temperature contrast, δT_b, relative to the CMB [10]. We created 6000 brightness temperature images [4] with resolution 1.5 Mpc through the semi-numerical code 21cmFast [9]. We varied two parameters corresponding to the cosmological context: the matter density parameter $\Omega_m \in [0.2, 0.4]$ and the rms linear fluctuation in the mass distribution on $8h^{-1}$Mpc $\sigma_8 \in [0.6, 0.8]$, and the other two parameters corresponding to the astrophysical context: the ionizing efficiency of high-z galaxies $\zeta \in [10, 100]$ and the minimum virial temperature of star-forming haloes $T^F_{vir} \in [3.98, 39.80] \times 10^4$K (hereafter represented in log10 units). For each set of parameters we produced 20 images at different redshifts in the range $z \in [6, 12]$, and stacked them into a single multi-channel tensor.

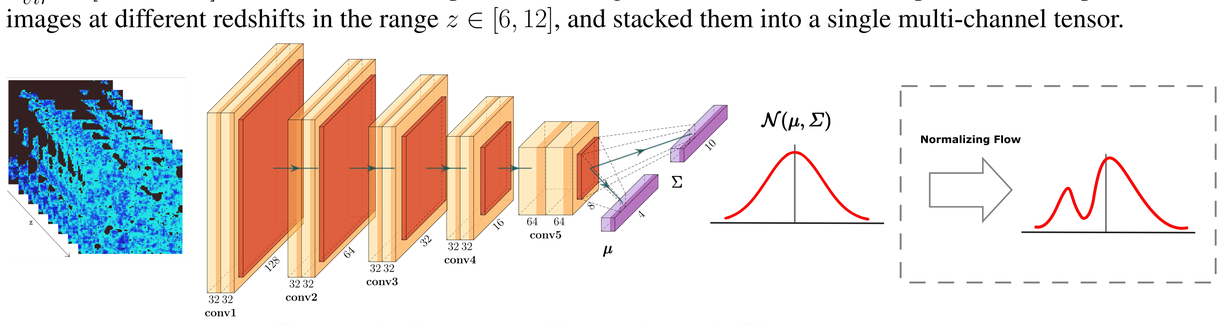

Figure 2: Bayesian Network with Flow in output.

We quantify the performance of the network by its coefficient of determination

$$R^2 = 1 - \frac{\sum_i (\bar{\boldsymbol{\mu}}(\boldsymbol{x}_i) - \boldsymbol{y}_i)^2}{\sum_i (\boldsymbol{y}_i - \bar{\boldsymbol{y}})^2} \tag{7}$$

and its confidence intervals, given by the percentage of times that $\boldsymbol{y}_i$ falls in a $\beta\%$ confidence interval, with $\beta = \{68.3, 95.5, 99.7\}$.

	Flipout (NLL=-2.4)				Dropout (NLL=-0.74)			
	σ_8	Ω_m	ζ	T^F_{vir}	σ_8	Ω_m	ζ	T^F_{vir}
R^2	0.92	0.97	0.83	0.97	0.87	0.94	0.65	0.92
C.L. 68.3%	74.1	70.2	75.4	70.3	70.4	67.3	58.5	76.1
C.L. 95.5%	97.4	96.2	98.5	96.0	95.7	96.3	91.7	98.5
C.L. 99.7%	99.7	99.9	99.9	99.9	99.6	99.8	99.8	99.9

We can observe that Flipout overperforms Dropout and gives more reliable uncertainty estimates. We can improve even more the accuracy of the errors by implementing post-processing calibration methods [4]. Flipout yields more accurate inferences and provides tighter constraints contours, see for example T^F_{vir}-ζ. Moreover, the correlations such as σ_8-Ω_m provide significant information for breaking parameter degeneracies and thus, be able to improve the existing measurements on cosmological parameters [8].

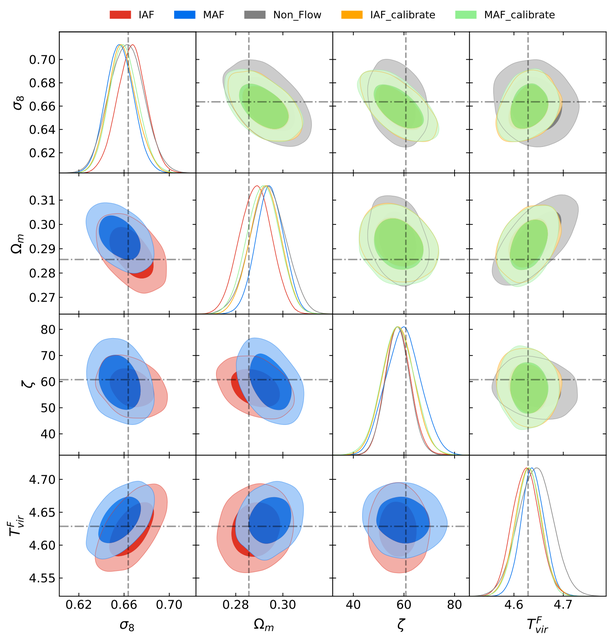

Figure 2: 68% and 95% contours from one example of our synthetic 21cm dataset.

	IAF (NLL=-3.80)			
	σ_8	Ω_m	ζ	T^F_{vir}
R^2	0.94	0.98	0.87	0.98
C.L. 68.3%	66.0	64.0	69.2	65.4
C.L. 95.5%	94.0	94.0	95.0	94.0
C.L. 99.7%	99.2	99.2	99.5	99.6

	MAF (NLL=-3.73)			
	σ_8	Ω_m	ζ	T^F_{vir}
R^2	0.94	0.98	0.87	0.98
C.L. 68.3%	64.7	63.7	69.1	65.0
C.L. 95.5%	93.3	94.2	95.1	94.0
C.L. 99.7%	99.0	99.3	99.3	99.4

	NVP (NLL=-3.44)			
	σ_8	Ω_m	ζ	T^F_{vir}
R^2	0.94	0.98	0.87	0.98
C.L. 68.3%	65.9	64.8	68.8	66.0
C.L. 95.5%	93.0	94.0	94.0	93.0
C.L. 99.7%	99.0	99.2	99.0	99.0

Figure 2: Metrics for the best experiments with Normalizing Flows after calibration.

3 The Cosmic Microwave Background

We generated 50.000 images related to the Cosmic Microwave Background (CMB) maps projected in $20 \times 20 deg^2$ patches in the sky using the script described in [5]. These images have size of (256,256,3) and each image corresponds to a specific set value of three parameters. We output a multivariate Gaussian, the NLL is

$$\mathcal{L} \sim \frac{1}{2}\log|\Sigma| + \frac{1}{2}(\mathbf{y}-\mu)^\top \Sigma^{-1}(\mathbf{y}-\mu)\ . \tag{8}$$

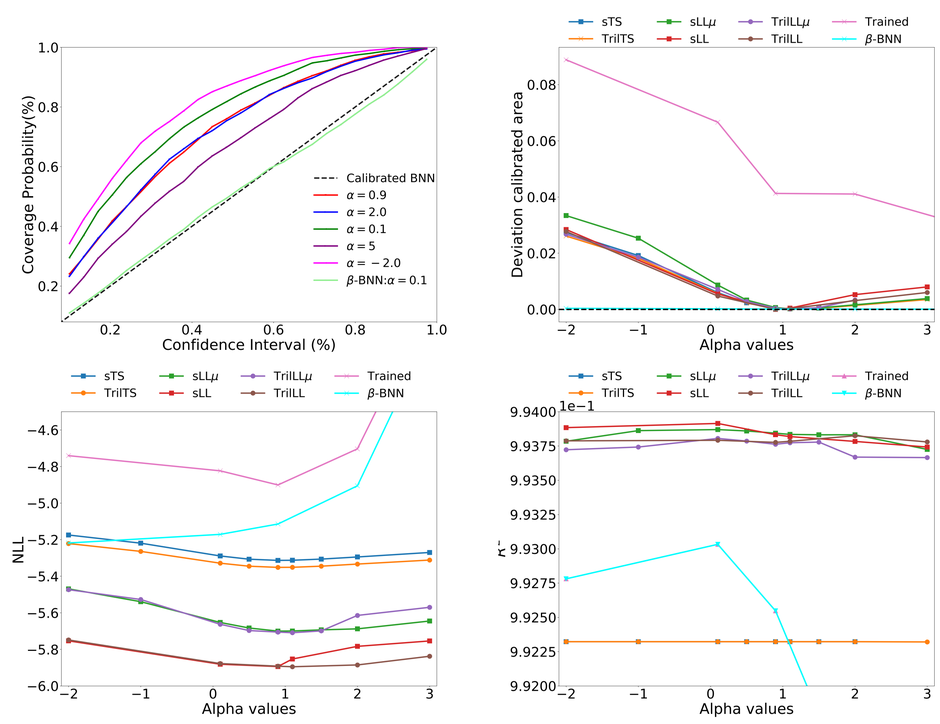

Figure 3: Reliability diagrams, miscalibration area, NLL and R^2 for varying α.

4 Adversarial Examples

We train a Classifier on the German traffic signs dataset GTSRB, with VGG blocks: 32, 64, 64, 128 and final linear layers: 512, 128, 43. Optionally we decide to preprocess the input images by reconstructing through a Variational Autoencoder (VAE) based on a residual encoder and decoders and a latent size of 100. We attack the network with Expectation over Transformation (EOT) with a number of samples of 10 and with a Carlini Wagner (CW) step. The attack is always performed in a white box scenario by attacking the full system, either Classifier or VAE+Classifier. We notice how BNNs classifiers trained with alphas 0.5, 0.9 and 1.1 provide a defense advantage with respect to the other Networks.

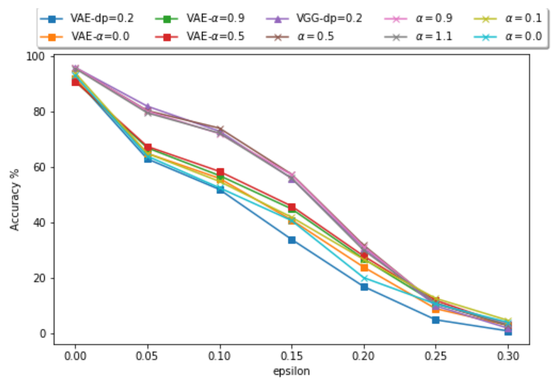

Figure 3: EOT CW attack for varying epsilon, and with different α during training.

These results show how a Bayesian approach, combined with the change in the optimization objective by varying alpha, enhance the robustness of the Classifier to adversarial examples. As a future development, we can calibrate the classifier after training it with KL, similarly to what we have done on the regression problems and compare the robustness of alpha in training vs alpha in calibration.

References

[1] S. Amari and H. Nagaoka. Translations of mathematical monographs, 2000.

[2] A. Graves. Practical variational inference for neural networks. NIPS, p.2348, 2011.

[3] J. M. Lobato, Y. Li, et al. Black-box α-divergence Minimization, Nov. 2015.

[4] H. J. Hortúa, R. Volpi, and L. Malagò. Parameters estimation from the 21 cm signal using variational inference. *Machine Learning: Science and Technology*, 2020.

[5] H. J. Hortua, R. Volpi, D. Marinelli, and L. Malagò. Parameters estimation for the cosmic microwave background with bayesian neural networks. *arXiv:1911.08508*, 2019.

[6] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? NIPS, p.5574, 2017.

[7] Y. Li and Y. Gal. Dropout inference in bayesian neural networks with alpha-divergences. ICML, p. 2052, 2017.

[8] M. McQuinn, O. Zahn, M. Zaldarriaga, L. Hernquist, and S. R. Furlanetto. Cosmological parameter estimation using 21 cm radiation from the epoch of reionization. *The Astrophysical Journal*, 653, 2006.

[9] A. Mesinger and S. Furlanetto. Efficient simulations of early structure formation and reionization. *The Astrophysical Journal*, 669, 2007.

[10] J. R. Pritchard and A. Loeb. 21 cm cosmology in the 21st century. *Reports on Progress in Physics*, 75, 2012.

[11] Y. Wen, P. Vicol, J. Ba, D. Tran, and R. Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. *arXiv:1803.04386*, 2018.