Biases in Variational Bayesian Neural Networks

Thang D. Bui
University of Sydney, Australia
thang.bui@sydney.edu.au

Abstract

Variational inference recently became the de facto standard method for approximate Bayesian neural networks. However, the standard mean-field approach (MFVI) possesses many undesirable behaviours. This short paper empirically investigates the variational biases [Turner and Sahani, 2011] of MFVI and other variational families. The preliminary results shed light on the poor performance of many variational approaches for model selection.

1 Introduction

Scalable stochastic variational inference methods have played a key role in the recent resurgence of Bayesian neural networks (BNNs) [see e.g. Hinton and van Camp, 1993, Graves, 2011, Blundell et al., 2015]. The most popular variational approach among all is one based on a mean-field (or diagonal) Gaussian variational approximation (MFVI). The MFVI variational objective bears similarity to a regularised loss function used in vanilla deep learning, hence is easy to implement and scale. Despite the popularity and successes, the approximation quality of MFVI for modern deep neural networks is still largely unknown. In many settings, MFVI has been shown to give pathological behaviours. For example, Trippe and Turner [2018] and Swaroop et al. [2019] demonstrate excessive hidden unit pruning when using MFVI on simple regression and classification tasks; Foong et al. [2019a,b] provably demonstrate mean-field variational BNNs with one hidden rectified linear layer unable to capture *in-between* uncertainty; and even more worryingly, Coker et al. [2021] show MFVI will ignore data in the infinite width limit.

In this short paper, we empirically investigate MFVI and other variational families using a slightly different lens, the marginal likelihood. In particular, we ask: how well do these variational approaches approximate the marginal likelihood, and if not very well, what does the variational bias [Turner and Sahani, 2011] looks like for variational BNNs?. The answers to these questions will shed light on the reliability of these variational approaches for hyper-parameter or model selection, and model comparison. We will first summarise the variational methods used for the comparison in section 2 and then present some preliminary results in section 3.

2 Variational inference for Bayesian neural networks

Consider a neural network with parameters $\mathbf{w}$ and a training dataset $\mathcal{D}$, exact inference and learning requires computing the posterior density, $p(\mathbf{w}|\mathcal{D},\theta)$, and the log marginal likelihood, $\mathcal{L}(\theta)$, as follows,

$$p(\mathbf{w}|\mathcal{D},\theta) = \frac{p(\mathbf{w}|\theta)p(\mathcal{D}|\mathbf{w},\theta)}{p(\mathcal{D}|\theta)}, \qquad \mathcal{L}(\theta) = \log p(\mathcal{D}|\theta) = \log \int_{\mathbf{w}} p(\mathcal{D}|\mathbf{w},\theta)p(\mathbf{w}|\theta), \quad (1)$$

where θ denotes the hyper-parameters, or architecture of the model, e.g. the prior variance or the observation noise variance in the regression setting. Whilst the posterior is useful at test time, the

Bayesian Deep Learning workshop, 35th Conference on Neural Information Processing Systems (NeurIPS 2021).

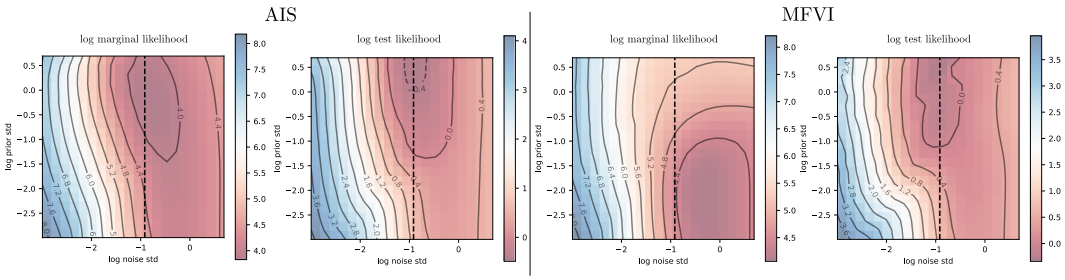

Figure 1: Negative log marginal likelihood estimate and negative test log likelihood (both in log scale) given by AIS and MFVI. Black dashed line indicates the true observation noise.

marginal likelihood is central to empirical Bayes approaches to model selection and comparison. However, both terms are intractable for all modern neural networks and thus require approximations. Variational inference conveniently provides a lower bound to the log marginal likelihood, which can be used for optimising a variational density *and*, in principle, selecting or comparing models,

$$\mathcal{F}(q(\mathbf{w}), \theta) = -\mathrm{KL}[q(\mathbf{w})||p(\mathbf{w}|\theta)] + \int_{\mathbf{w}} q(\mathbf{w}) \log p(\mathcal{D}|\mathbf{w}, \theta). \tag{2}$$

The variational families (aka. the form of $q(\mathbf{w})$) governs the flexibility of the approximate posterior and thus the tightness of the variational lower bound. We consider the following variational approximations:

- mean-field Gaussian variational inference (MFVI) with $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{m}, \mathrm{diag}(\mathbf{v}))$, where $\mathbf{m}$ and $\mathbf{v}$ represent the means and variances of network weights.
- fully-correlated Gaussian variational inference (FCVI) with $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{m}, \mathbf{V}))$, where $\mathbf{m}$ and $\mathbf{V}$ represent the mean and covariance of network weights. Note that this approach is only tractable for small networks.
- importance weighted variational inference (IWVI) [Burda et al., 2015] with a mean-field Gaussian proposal as in MFVI.
- VI using the Thermodynamic Variational Objective (TVO) [Masrani et al., 2019]. The key challenge for both IWVI and TVO for BNNs is the need to sample from a high-dimensional and potentially inaccurate approximate posterior, leading to high variance.
- an autoregressive, conditionally Gaussian approximation based on inducing points (IPVI) [Ober and Aitchison, 2021]. This approximation assumes a correlated Gaussian density for the first layer's weights, but all other layers' are marginally non-Gaussian and that all weights in the network are correlated.

We also consider an approximation based on α-divergence (BB-α) [Hernandez-Lobato et al., 2016], with a mean-field Gaussian approximation as in MFVI. As some of the methods above do not admit unbiased gradients using data minibatches, we opt for a comparison using the full dataset at each gradient step. We use 5 samples from the approximate posterior or proposal for all methods, train each method for 5000 epochs, and use 50 log-spaced intermediate distributions for TVO.

Note that all the methods above are approximate and thus only give an approximation to the log marginal likelihood. To assess the fidelity of these approximations, we run annealed importance sampling (AIS) [Neal, 2001] with 5000 intermediate distributions and 100 parallel chains, and hybrid Monte Carlo (HMC) [Neal, 2011] with 10 leapfrog steps to sample from the intermediate distributions for the regression datasets considered in section 3. AIS requires the full dataset at each HMC gradient step and a large number of intermediate densities, and is, therefore, not tractable for big networks and large datasets. We only use it here for evaluation, in a similar spirit to how Izmailov et al. [2021] use HMC to evaluate approximate inference methods.

3 Experimental results

In this section, we summarise the results of the aforementioned methods on a 1D synthetic regression dataset and a neural network with 50 ReLUs. The observations on other regression/classification

datasets and when using tanh units are almost identical. Similar to [Turner and Sahani, 2011], we vary the hyperparameters in the model, prior variance and observation noise variance and evaluate the log marginal likelihood approximation after training each method. We also evaluate the performance of these methods on a test set. All results are averaged using five random seeds.

3.1 Biases of mean-field variational inference

We first plot the log marginal likelihood and test log likelihood surfaces when varying the hyperparameters in fig. 1. It could be seen that (i) AIS marginal likelihood and its test performance strongly correlate, (ii) MFVI marginal likelihood is biased towards larger noise and smaller prior variances, and (iii) critically, MFVI test performance is not correlated with MFVI marginal likelihood. The last two points explain the use of a validation set, instead of the variational bound, to select hyperparameters or models in practice [see e.g. Blundell et al., 2015]. To understand these issues further, we plot the MFVI and HMC predictions using the best hyperparameters suggested by AIS and MFVI in fig. 2. Due to the mean-field assumption and the KL term in the variational bound, MFVI *prefers* a large observation noise and forces all network weights to match the prior.

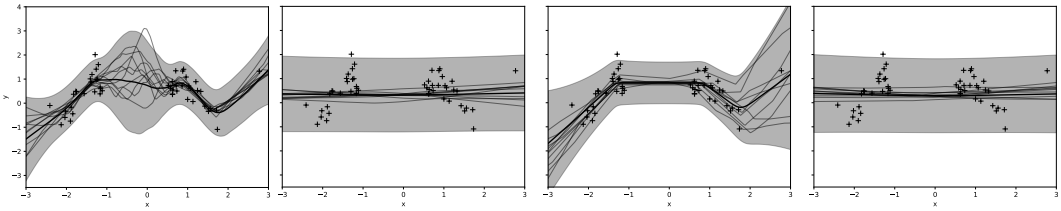

Figure 2: Training data and predictions given by AIS and MFVI at hyperparameters selected using the marginal likelihood in fig. 1. Left to right: HMC prediction using the best AIS hypers, HMC prediction using the best MFVI hypers, MFVI prediction using the best AIS hypers, and MFVI prediction using the best MFVI hypers. Due to the variational biases, MFVI suggests the model in the fourth plot has a higher log marginal likelihood than that of the model in the third plot.

3.2 Biases of more structured approximations and other divergences

We repeat the same experiment as above using other variational families, and plot the log marginal likelihood estimates and test performance in fig. 3. Note that all other mean-field variational methods exhibit the same pathological behaviour as MFVI. FVCI, initialised at the MFVI solution, does not seem able to escape from the MFVI local optimum. IPVI surprisingly outperforms all other methods and closely tracks the AIS estimate. This suggests structured, non-Gaussian variational approximations are crucial to obtain accurate log marginal likelihood estimates.

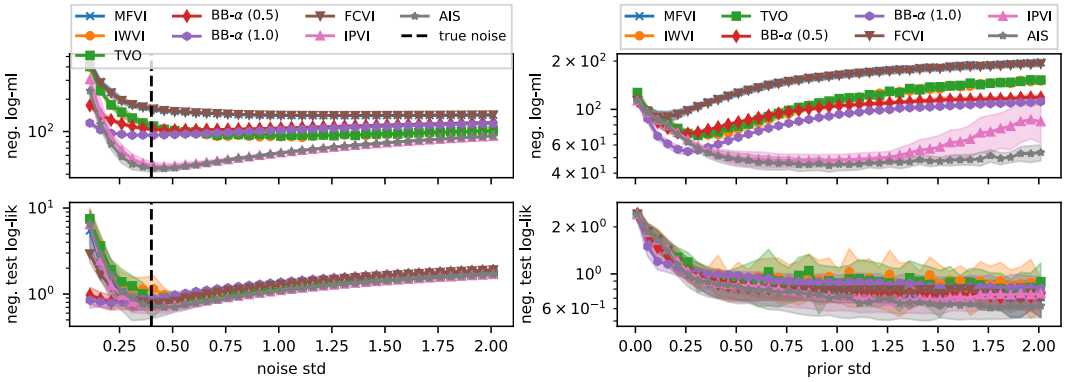

Figure 3: Marginal likelihood estimates and test likelihood provide by AIS and several variational approximations. Left: the prior std is fixed to 1. Right: the observation noise std is fixed to the true noise level, 0.4.

4 Summary

We have empirically shown the hyper-parameter biases of several variational methods on Bayesian neural networks. The results, despite preliminary, demonstrate the benefit of more structured, non-Gaussian variational approximations, and suggest a direction of research into empirical Bayes using AIS or by direct estimation of the log marginal likelihood gradients [Tomczak and Turner, 2021].

References

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1613–1622, 2015.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Beau Coker, Weiwei Pan, and Finale Doshi-Velez. Wide mean-field variational Bayesian neural networks ignore the data, 2021.

Andrew Y. K. Foong, Yingzhen Li, José Miguel Hernández-Lobato, and Richard E. Turner. 'in-between' uncertainty in Bayesian neural networks, 2019a.

Andrew YK Foong, David R Burt, Yingzhen Li, and Richard E Turner. On the expressiveness of approximate inference in Bayesian neural networks. *arXiv preprint arXiv:1909.00719*, 2019b.

Alex Graves. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Jose Hernandez-Lobato, Yingzhen Li, Mark Rowland, Thang Bui, Daniel Hernandez-Lobato, and Richard Turner. Black-box alpha divergence minimization. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1511–1520, 2016.

Geoffrey E. Hinton and Drew van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, page 5–13, 1993.

Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. What are Bayesian neural network posteriors really like? In *Proceedings of the 38th International Conference on Machine Learning*, pages 4629–4640, 2021.

Vaden Masrani, Tuan Anh Le, and Frank Wood. The thermodynamic variational objective. In *Advances in Neural Information Processing Systems*, 2019.

Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.

Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, page 113, 2011.

Sebastian W Ober and Laurence Aitchison. Global inducing point variational posteriors for Bayesian neural networks and deep Gaussian processes. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8248–8259, 2021.

Siddharth Swaroop, Cuong V. Nguyen, Thang D. Bui, and Richard E. Turner. Improving and understanding variational continual learning, 2019.

Marcin B. Tomczak and Richard Turner. Marginal likelihood gradient for Bayesian neural networks. *AABI*, 2021.

Brian Trippe and Richard Turner. Overpruning in variational Bayesian neural networks. *arXiv preprint arXiv:1801.06230*, 2018.

R. E. Turner and M. Sahani. Two problems with variational expectation maximisation for time-series models. In D. Barber, T. Cemgil, and S. Chiappa, editors, *Bayesian Time series models*, chapter 5, pages 109–130. Cambridge University Press, 2011.