Diversity is All You Need to Improve Bayesian Model Averaging

Yashvir Grewal
Monash University, Australia
yashvirgrewal@gmail.com

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

Abstract

Existing approximate inference techniques produce predictive distributions that are quite distinct from the predictive distribution of the gold-standard Hamiltonian Monte Carlo. In this work, we bring the predictive distribution produced by deep ensembles more closer to the Hamiltonian Monte Carlo predictive distribution by increasing the diversity within the ensembles. The proposed approach outperforms the existing approximate inference methods and is also currently ranked the highest in the Approximate Inference competition at NeurIPS 2021.

1 Introduction

The synthesis of deep neural networks and probabilistic modelling forms a promising path to obtain calibrated uncertainty estimates, which is vital in many real-world decision-making applications. The wrinkle is, however, the exact posterior for all modern neural networks is analytically and computationally intractable. A plethora of approximate methods have thus been developed in the last few years, allowing Bayesian deep learning to be applied at scale [see e.g. Welling and Teh, 2011, Blundell et al., 2015, Gal and Ghahramani, 2016, Lakshminarayanan et al., 2017, Osawa et al., 2019]. However, the quality of these approximations has recently come under the spotlight. For example, these methods can perform poorly in uncertainty-driven sequential decision making [Riquelme et al., 2018] or when dealing with out-of-distributions inputs [Ovadia et al., 2019]. More crucially, Izmailov et al. [2021] show that the predictive distributions learnt by these methods are dissimilar to that produced by the gold-standard Hamiltonian Monte Carlo (HMC) [Neal, 2011], a non-scalable method that asymptotically samples from the true posterior.

In this work, we revisit the Bayesian model averaging approximation using deep ensembles [Lakshminarayanan et al., 2017], and proposes two simple, yet effective modifications: (i) using different optimisers for ensemble members, and (ii) combining predictions from multiple network architectures on the same dataset. The proposed approach is simple to use, just like deep ensembles, but yields predictive distributions that resemble the HMC predictive. This is highlighted by a strong improvement over the baseline methods along with currently the highest ranked scores in the Approximate Inference in Bayesian Deep Learning competition at NeurIPS 2021.

2 Proposed approach

We start off with formalising the target predictive distribution as a Bayesian model average (BMA) and its approximation using deep ensembles. The target distribution can be written as:

$$p(y|x, \mathcal{D}) = \int_{\mathbf{w}} p(y|x, \mathbf{w}) p(\mathbf{w}|\mathcal{D}) \tag{1}$$

where $\mathcal{D}$ is the training dataset and $\mathbf{w}$ is the parameters in a neural network. Wilson and Izmailov [2020] argue that deep ensembles [Lakshminarayanan et al., 2017] is a valid BMA approximation,

Preprint. Under review.

in the same vein as other approximate Bayesian inference methods. In particular, deep ensembles trains a single neural network several times with different initial weights. Because of a distinct initialisation and stochasticity in the training procedure, each member within the ensemble can end up in a different basin of the posterior. The weights sampled from these distinct basins tend to give diverse predictions, thus yielding an arguably better BMA approximation than other single-mode methods [Fort et al., 2020, Wilson and Izmailov, 2020].

In practice, the optimisation procedure of deep ensembles involves using the same optimisation algorithm to train each member. However, Zhou et al. [2020] show that optimisation algorithms such as SGD and Adam have an implicit attraction towards certain basins with specific geometric properties. For example, SGD has an affinity towards flat basins, while Adam has a bias towards sharper basins. This means that if all members of the ensemble are trained with the same optimizer, we may miss converging on some basins because of their geometric properties. This would prevent the BMA approximation to bias away from certain regions. Even though flat regions may support better generalisation and occupy a larger posterior volume, it is worth considering other basins in the BMA marginalisation. To this end, we explore using different optimisation algorithms within the ensemble. Similar to deep ensembles, we train each network in the ensemble with a different initialisation of the weights. However, instead of training each network with the same optimiser like in deep ensembles, we train half of the networks with SGD and the other half with Adam.

Note that BMA in eq. (1) can be extended, in principle, to combine predictions from multiple models,

$$p(y|x, \mathcal{D}) = \sum_{\mathbf{M} \in \mathcal{M}} \int_{\mathbf{w}} p(y|x, \mathbf{w}, \mathbf{M}) p(\mathbf{w}|\mathbf{M}, \mathcal{D}) p(\mathbf{M}|\mathcal{D}), \tag{2}$$

that is BMA linearly mixes the predictions from multiple models, each weighted by its posterior probability. Even though this seems appealing, there are several subtleties. First, eq. (2) can be pathological when considered models do not capture mutually exclusive and exhaustive possibilities about how the data was generated [Minka, 2002]. Second, it is intractable to consider all models or network architectures, so we have to resort to an approximate using a finite number. And, third, even when the model space is small, getting the posterior probability for each model is intractable. Nevertheless, we may still wish to combine the predictions from multiple models. In this work, we simply train and combine predictions from several networks with well-known architectures for a given dataset. In the same spirit as deep ensembles having several initialisations or optimisers, including different architectures in the BMA tends to give more functional diversity. As we show in our experiments, this increased functional diversity brings the predictive distribution much more closer to the exhaustive HMC in comparison to the existing methods.

3 Experiments

To evaluate our approach, we follow the evaluation framework used by Izmailov et al. [2021]. We use the total variation and agreement to measure how close the predictive distribution of a method is to that of HMC. The total variation (lower is better) compares the probabilities for each of the classes between the predictive distribution p and the true distribution q. Total variation is defined as:

$$\frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \sum_{j} |p(y = j|x_i) - q(y = j|x_i)| \tag{3}$$

where x_i is the ith data point and n is the total number of data points. Agreement (higher is better) computes the similarity of top-1 predictions of p and q and is defined as

$$\frac{1}{n} \sum_{i=1}^{n} I[\arg\max_j p(y = j|x_i) = \arg\max_j q(y = j|x_i)] \tag{4}$$

where I is the indicator function.

Within our ensemble, we include equal number of networks trained using both SGD and Adam. To take into account predictions from different models, our ensemble also consists of equal number of

Table 1: Agreement and total variation of the predictive distribution obtained by our approach and existing approximate inference methods with HMC on CIFAR-10. The HMC reference distribution here represents the agreement and variation of a single HMC chain with an ensemble of two independent HMC chains. Our proposed approach, which includes varying both the optimizer and the network architecture in the ensemble (Opt and Model Ens), outperforms all the existing baseline methods. For comparison, we also include an ensemble in which we only vary the optimizer and leave the architecture fixed (Opt only Ens).

Metric	HMC (Reference)	**Opt and Model Ens**	Opt only Ens	Deep Ens	MVFI	SGLD	SGHMC CLR	SGHMC CLR-Prec
Acc	89.64	**90.19**	89.56	88.49	86.45	89.32	89.63	87.46
Agreem	94.01	**93.79**	92.35	91.52	88.75	91.54	92.67	90.96
Total Var	0.074	**0.087**	0.102	0.115	0.136	0.110	0.099	0.111

Resnet-20 [He et al., 2016] and Lenet-5 models [Lecun et al., 1998], each trained with both SGD and Adam. The ensemble consists of 16 networks in total. We report the results on CIFAR-10 in Table 1. We use the baseline results of the existing methods provided by Izmailov et al. [2021] for comparison.

Our proposed approach outperforms all the existing methods in terms of both total variance and agreement with the HMC distribution. It also has a higher final accuracy than a single chain HMC as well as all other methods. It is also worth noting that our ensemble approach is relatively much more efficient as it only consists of 16 networks in comparison to the ensemble of 50 networks used by Izmailov et al. [2021], reported as deep ensembles in table 1. In addition to CIFAR-10, we also test our approach on IMDB dataset, where we use an ensemble of Adam and RMSProp with CNN-LSTM and Bi-directional LSTM. The agreement and total variance with HMC we achieve are also currently ranked the highest in the Approximate Inference Competition.

4 Summary and future work

Using multiple optimisers and models within a deep ensemble enables greater functional diversity, leading to promising results on two large scale approximate inference settings. There is still much work to be done on understanding these modifications and their impact on other downstream applications such as distribution shift detection [Ovadia et al., 2019].

Deep ensembles [Lakshminarayanan et al., 2017] can be made efficient by sharing parameters between ensemble members [Wen et al., 2020]. In addition, the functional diversity can be further improved by using random hyperparameters for different members [Wenzel et al., 2020]. The proposed approach here is compatible with these existing methods and we plan to explore the combination of these in future work.

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.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective, 2020.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

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.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Thomas P Minka. Bayesian model averaging is not model combination. 2002.

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

Kazuki Osawa, Siddharth Swaroop, Mohammad Emtiyaz E Khan, Anirudh Jain, Runa Eschenhagen, Richard E Turner, and Rio Yokota. Practical deep learning with Bayesian principles. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Carlos Riquelme, George Tucker, and Jasper Roland Snoek. Deep Bayesian bandits showdown. 2018. URL `https://openreview.net/pdf?id=SyYe6k-CW`.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, 2011.

Yeming Wen, Dustin Tran, and Jimmy Ba. BatchEnsemble: An alternative approach to efficient ensemble and lifelong learning, 2020.

Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. Hyperparameter ensembles for robustness and uncertainty quantification. In *Advances in Neural Information Processing Systems*, volume 33, pages 6514–6527, 2020.

Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization, 2020.

Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Hoi, and Weinan E. Towards theoretically understanding why SGD generalizes better than ADAM in deep learning, 2020.