International Statistical Review

International Statistical Review (2020), 0, 0, 1–3 doi:10.1111/insr.12377

Discussion

Soumya Ghosh[1] and Finale Doshi-Velez[2]

[1] *MIT-IBM Watson AI Lab and Center for Computational Health, IBM Research, Cambridge, MA, USA*

[2] *School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA*
E-mail: ghoshso@us.ibm.com)

We congratulate the authors on a timely survey of horseshoe priors and their applications. Horseshoe priors have been analysed and applied in so many communities; one of the key contributions of this survey is that it succinctly reviews and synthesises our statistical understanding of these priors to date. By highlighting work on models employing non-Gaussian noise, nonlinear, and compositional structure, the survey underscores the utility of the horseshoe distribution in contexts well beyond its traditional application in linear models with Gaussian noise.

The authors also review various computational approaches that have been developed to perform approximate inference under this model. In this discussion, we take the opportunity to further elaborate on the many computational issues faced in working with such priors, especially for large, overparameterised models trained with large amounts of data. A Bayesian neural network (BNN) is a canonical example of models of this class. Beyond identifiability, overparameterisation can cause BNNs to fit the data poorly (see illustration in ; Ghosh *et al.*, 2019). The strong shrinkage towards zero provided by the horseshoe and related global–local priors are particularly attractive for BNNs: by identifying a subset of unshrunk parameters and strongly shrinking others, they alleviate issues stemming from overparameterisation.

However, in addition to having many parameters, these neural models are typically used in settings where the datasets are quite large, making approximate Bayesian inference challenging. While Hamiltonian Monte Carlo (HMC), an MCMC algorithm that exploits gradient information to propose likely configuration of parameters, remains the gold standard for posterior inference, it requires a Metropolis–Hastings correction to decide whether to accept a proposal. Both the statistic required for the correction and the gradient required for the proposal depend on all data and are prohibitively expensive to compute for typical datasets used for training neural networks.

As a result, much recent effort has focused on developing variational approximations to the intractable posterior. Variational inference approximates an intractable posterior $p(\theta \mid \{x_i, y_i\}_{i=1}^{n})$ with a tractable alternative $q(\theta \mid \phi)$ by minimising the Kullback–Leibler divergence, $\mathrm{KL}(q(\theta \mid \phi)||p(\theta \mid \{x_i, y_i\}_{i=1}^{n}))$ with respect to the variational free parameters, ϕ. This is equivalent to minimising the *variational free energy*, $\mathcal{F}(\phi)$,

$$\mathcal{F}(\phi) = -\mathbb{E}_q(\theta \mid \phi)\ln p(\theta) + \ln \sum_{i=1}^{n} p(y_i \mid x_i, \theta) - \mathbb{H}[q(\theta \mid \phi)], \qquad (1)$$

where $\mathbb{H}[q(\theta \mid \phi)]$ denotes the entropy of the distribution, $q(\theta \mid \phi)$. For large n, one can obtain a noisy but unbiased estimate of the gradient of the variational free energy $\mathcal{F}(\phi)$ by replacing the sum over all n data points with a sum over a smaller number of indices and appropriately scaling the prior and entropy terms (Hoffman *et al.*, 2013). In the case of BNNs, the expectations in Equation (1) have no closed form; in practice, these are approximated with Monte Carlo expectations. Given unbiased gradients, stochastic gradient descent approaches are invoked to minimise $\mathcal{F}(\phi)$. The full procedure is sometimes referred to as being *doubly stochastic* owing to the two sources of stochasticity—data subsampling and the Monte Carlo expectation estimate—and is black box in the sense that it is largely agnostic about the particulars of the model $p(x, y, \theta)$.

Doubly stochastic variational inference has been popular for inference in BNNs(Blundell *et al.*, 2015) and has also been used for learning BNNs with horseshoe priors(Ghosh & Doshi-Velez, 2017; Ghosh *et al.*, 2019). The so-called "reparameterization trick" employed in these papers refers to a particular procedure for computing Monte Carlo gradients of $\mathcal{F}(\phi)$(Mohamed *et al.*, 2019). For horseshoe BNNs, we are interested in the posterior over the network weights and their respective scales. These parameters tend to be strongly correlated, with small scales resulting in strongly shrunk weights while larger scales allow the corresponding weights to escape unshrunk. These interactions induce challenging posterior geometries that are difficult to reliably sample or approximate. Adopting alternate noncentred parameterisations of the model help alleviate some of these difficulties. The reparameterisation trick and the noncentred parameterisation constitute two orthogonal improvements—one allows for scalable inference by differentiating through Monte Carlo samples while the other improves the quality of the inference.

Different choices of variational families provide interesting trade-offs between modelling the interactions that occur in horseshoe BNN posteriors and computation. Restricting the variational posterior over these quantities to Gaussian and log-normal families while ignoring correlations between them allows us to develop a computationally convenient instance of doubly stochastic variational inference. One can further reduce computation at the expense of accuracy by fixing the variances and optimising only the means of the variational Gaussians over weights (as done in ; Ghosh & Doshi-Velez, 2017). It is also possible to retain more of the posterior structure in the variational approximation. For example, in Ghosh *et al.* (2019), we found that modelling correlations among weights in a network's layer as well as between weights and their respective scales provided both stronger shrinkage and better calibrated predictions, at the expense of increased computation.

One could imagine exploring variational approximations at more extreme ends of the computation-accuracy spectrum. For instance, a fully factorised approximation that employs fixed variance distributions over *both* weights and scales could be used. However, we do not find such approximations very attractive. First, it is common to place (group) horseshoe priors over network units (see ; Ghosh & Doshi-Velez, 2017; Ghosh *et al.*, 2019) rather than over individual weights. The scale parameters thus only grow linearly with the width of a layer; tying parameters of their variational approximations provide only a modest computational benefit. Moreover, empirically we find such approximations to be quite inaccurate. At the other end of the spectrum, one could use approximating families that do not factorise even across layers in a network. While jointly modelling the weight parameters is computationally infeasible for all but the smallest networks, using a distribution over scales that does not factorise within or across layers may be feasible. It remains to be seen whether such elaborate variational families indeed yield better approximate posteriors or succumb to optimisation challenges.

Apart from these trade-offs, moving beyond unimodal Gaussian approximations, exploring alternate auxiliary variable representations of the horseshoe distribution, and developing algorithms for capturing posterior multimodality are all likely to be useful for better characterising the posterior. Recent advances in stochastic MCMC algorithms(Ma *et al.*, 2015) constitute another promising direction, but would need to overcome difficulties stemming from challenging posterior geometries exhibited by the horseshoe distribution and multimodality exhibited by BNNs.

References

Blundell, C., Cornebise, J., Kavukcuoglu, K. & Wierstra, D. (2015). Weight Uncertainty in Neural Networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. Lille, France, pp. 1613–1622.

Ghosh, S. & Doshi-Velez, F. (2017). Model Selection in Bayesian Neural Networks via Horseshoe Priors. *NIPS Work. Bay. D. Learn.*

Ghosh, S., Jiayu, Y. & Finale D.-V. (2019). Model selection in Bayesian neural networks via horseshoe priors. *J. Mach. Learn. Res.*, **20**(182), 1-46.

Hoffman, M.D., Blei, D.M., Wang, C. & Paisley, J. (2013). Stochastic variational inference. *J. Mach. Learn. Res.*, **14**(1), 1303–1347.

Ma, Y.-A., Chen, T. & Fox, E. (2015). A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*. Montreal, Quebec, Canada, pp. 2917–2925.

Mohamed, S., Rosca, M., Figurnov, M. & Mnih, A. (2019). Monte Carlo gradient estimation in machine learning. *arXiv preprint arXiv:1906.10652*.

[*Received January 2020, accepted March 2020*]