What Are Bayesian Neural Network Posteriors Really Like?

Pavel Izmailov Sharad Vikram Matthew D. Hoffman Andrew Gordon Wilson

NYU Google AI

Overview

We run high-fidelity HMC on hundreds of TPU devices for millions of training epochs to provide our best approximation of true Bayesian neural networks (BNNs).

- BNNs outperform deep ensembles
- No cold posteriors needed for good performance
- Deep ensembles more like HMC than mean-field variational inference
- BNNs are surprisingly poor under data corruption
- Parameter-space priors have a limited effect, Gaussians perform well

Bayesian neural networks

Bayesian inference is especially compelling for deep neural networks!

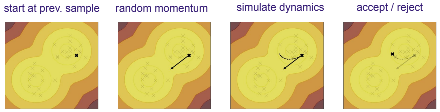
$$p(w|\text{Data}) \propto p(\text{Data}|w) \cdot p(w)$$

$$p_{BMA}(y|x) = \int p(y|w,x)p(w|\text{Data})dw \approx \frac{1}{n}\sum_i p(y|w_i,x)$$

$$w_i \sim p(w|\text{Data})$$

HMC: Hamiltonian Monte Carlo

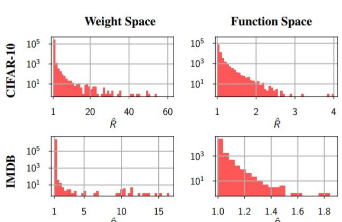
Simulating the dynamics of a particle sliding on the plot of the log-density function that we are trying to sample from



+ Asymptotically exact
+ Well-studied and understood
+ Has been used in early BNNs

- Requires exact gradients
- Generally expensive

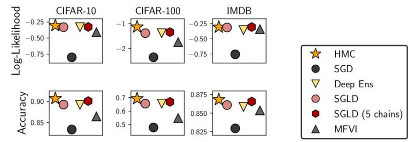
How well is HMC mixing?



$$\hat{R} \approx \frac{\text{between-chain variance}}{\text{avg within-chain variance}}$$

Most function space R are close to 1, indicating good mixing in function space.

Do BNNs perform well in practice?

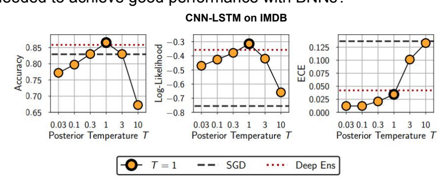


Bayesian neural networks achieve strong results outperforming even large deep ensembles.

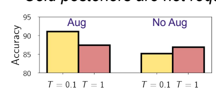
Do we need cold posteriors?

$$p_T(w|\mathcal{D}) \propto (p(\mathcal{D}|w) \cdot p(w))^{1/T}$$

Cold posteriors effect [Wenzel 2020]: cold posteriors (T << 1) are needed to achieve good performance with BNNs?

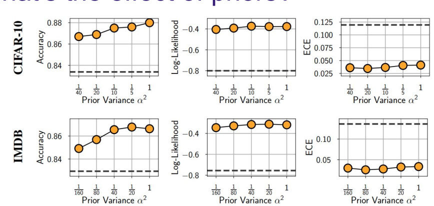


Cold posteriors are not required for good results and in fact can hurt!



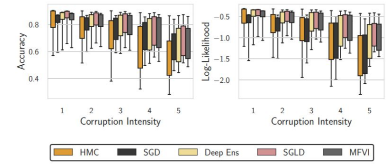
In fact, if we disable data augmentation in the code of [Wenzel 2020], there is no cold posteriors effect.

What's the effect of priors?

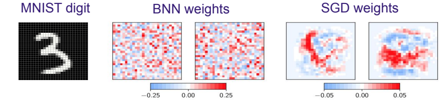


Results are fairly robust across $\mathcal{N}(0, \alpha^2 I)$ prior scale α as well as across prior families. The architecture dominates in prior specification.

Are BNNs robust to covariate shift?



HMC BNNs are terrible on corrupted data!



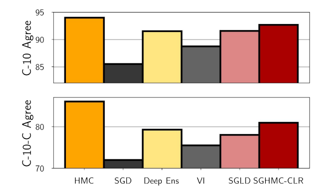
In [Izmailov 2021] we explain this phenomenon and provide a remedy.

How close are other methods to HMC?

All scalable BDL methods make distinct predictions from HMC.

SGMCMC provide the closest results.

Deep ensembles are not a "non-Bayesian competitor" to scalable BDL methods.



Deep ensembles are closer to HMC than mean-field variational inference!

References

[Wenzel 2020]: *How Good is the Bayes Posterior in Deep Neural Networks Really?*; Wenzel et al., ICML 2020

[Izmailov 2021]: *Dangers of Bayesian Model Averaging under Covariate Shift*; Izmailov, Nicholson, Lotfi, Wilson

Paper Code Samples Competition