Scaling Hamiltonian Monte Carlo Inference for Bayesian Neural Networks with Symmetric Splitting Supplementary Material

Adam D. Cobb[1] **Brian Jalaian**[1]

[1]US Army Research Laboratory, Adelphi, Maryland, USA

A VEHICLE CLASSIFICATION FROM ACOUSTIC SENSORS

A.1 DATA

In this section, we provide further details of the data set. Figure 1 shows what the input domain looks like and Figure 2 is a histogram showing the total data distribution.

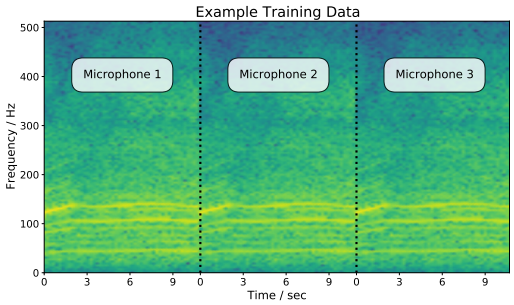


Figure 1: An example of a single input datum. The spectrograms from all three microphones (aligned in time) are concatenated into one image which is then passed into the CNN. The total 129×150 array has a resolution of 4.0 Hz in the vertical axis and a resolution of 0.22 seconds in the horizontal axis.

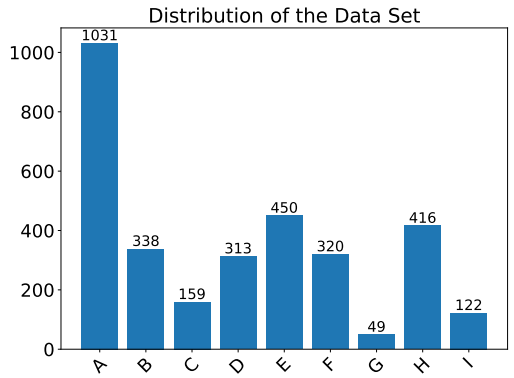


Figure 2: Histogram showing the distribution of the data set. Notice the large data imbalance, especially when comparing vehicle class 'G' to vehicle class 'A'.

Supplement for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021).

A.2 HYPERPARAMETER OPTIMISATION

The hyperparameters of all approaches were found via Bayesian optimisation (BO). For symmetric split HMC, we performed BO over vanilla HMC with a smaller subset of the data to reduce the computation time.

Stochastic Gradient Descent: Learning rate $= 0.0103$; momentum $= 0.9$; epochs $= 209$; weight decay $= 0.0401$; batch size $= 512$.

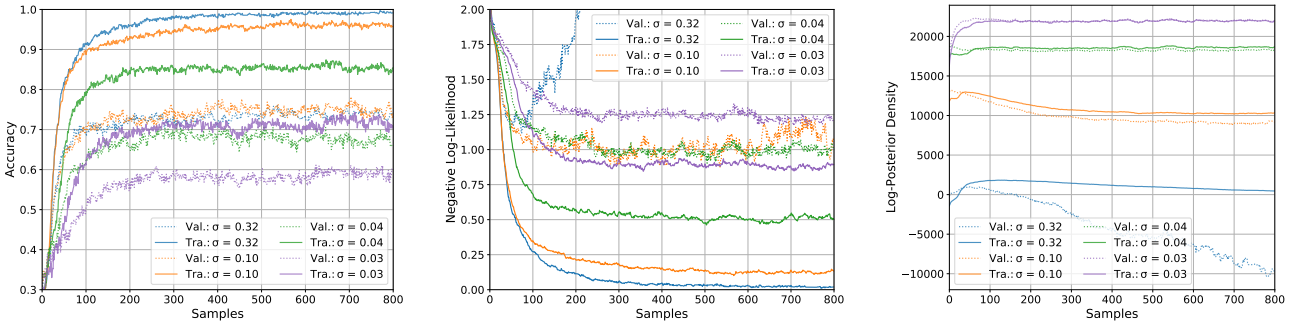
SGLD: Learning rate $= 0.0182$; prior standard deviation $= 0.7431$; epochs $= 400$; batch size $= 512$; burn $= 200$.

Stochastic Gradient HMC: Learning rate $= 0.0076$; prior standard deviation $= 0.1086$; epochs $= 1850$; friction term $= 0.01$; batch size $= 512$; burn $= 150$.

Symmetric Split HMC: $L = 11$; $\epsilon = 4.96e^{-6}$; $\mathbf{M} = 2e^{-5}\mathbf{I}$; $p(\boldsymbol{\omega}) = \mathcal{N}(\mathbf{0}, \tau^{-1}\mathbf{I})$, (with $\tau = 100$); number of splits $= 2$ (each of batch size 939); number of samples $= 3000$; burn $= 300$.

A.3 EFFECT OF THE PRIOR

We use this section as an opportunity to demonstrate the effect of the prior on the classification results. Each weight in our network has a univariate Gaussian prior with a variance of σ^2 (i.e. $p(\boldsymbol{\omega}) = \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$). We perform four experiments over the acoustic vehicle classification data, where $\sigma = 0.32, 0.10, 0.04$, and 0.03 are used for each implementation.[1] Figure 3 shows the importance of carefully selecting the prior. Setting σ to the larger (more flexible) value of 0.32 leads to over-fitting. For example in Figure 3a, the solid blue curve yields near-perfect accuracy over the training data, with the validation curve also displaying a good accuracy performance. However this model is misspecified, which can easily be seen from the validation Negative Log-Likelihood (NLL) performance of Figure 3b, which rapidly increases after approximately 100 samples (see dotted blue curve). This misspecification is especially obvious, when we plot the Log-Posterior Density in Figure 3c. Unlike the accuracy and the NLL, the Log-Posterior Density indicates the model is performing poorly from simply observing the performance over training data, where we see the solid blue curve continuing to decrease with the number of samples (and not stabilising at a value like in the other settings). These three indicators are especially important, as simply relying on accuracy would make it hard to distinguish between the best and worst performing models.



(a) Accuracy (b) Mean Negative Log-Likelihood (c) Log-Posterior Density

Figure 3: A performance comparison by varying the strength of the prior over the acoustic vehicle classification data set. We show the (a) Accuracy, (b) Negative Log-Likelihood (NLL), and (c) Log-Posterior Density. The curves are shown for $\sigma = 0.32, 0.10, 0.04$, and 0.03, where the prior $p(\boldsymbol{\omega}) = \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. When σ is too large, the samples achieve almost 100% accuracy and zero NLL in the training data. However this is at the severe cost of the validation performance (see blue curves). When σ is too small, the strength of the prior prevents the model from fitting to the data, with a lower accuracy and a higher NLL (see purple curves). The Log-Posterior Density acts as a good proxy for indicating convergence, by displaying the over-fitting collapse of $\sigma = 0.32$, where both the training and validation curves do not plateau.

[1] Corresponding to precisions of $10, 100, 500$, and 1000 respectively.

A.4 CLASSIFICATION PERFORMANCE: SUPPLEMENTARY PLOTS

We can also look at how the performance of the accuracy changes with the number of samples. Figure 4 shows the cumulative accuracy as the number of samples increases. In particular, we show the last $1{,}500$ samples for all MCMC schemes. Note that SGLD and SGHMC plateau at a lower mean accuracy.

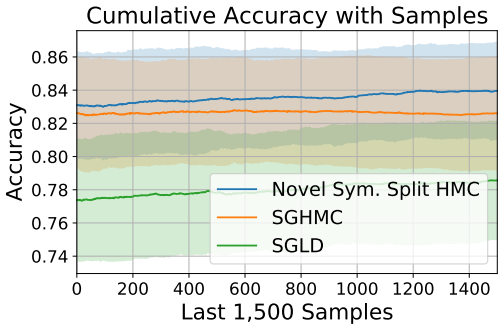


Figure 4: Cumulative accuracy of the ensemble of model samples. The standard deviation is over the cross-validation splits. The accuracy at each step is calculated by comparing the true label with $\max_c \mathbb{E}_{\boldsymbol{\omega}}[p(\mathbf{y}^* = c|\mathbf{x}^*)]$, where the expectation is over the samples up until that point. Symmetric split HMC continues to improve with the materialised number of samples.