Piecewise Deterministic Markov Processes for Bayesian Neural Networks (Supplementary Material)

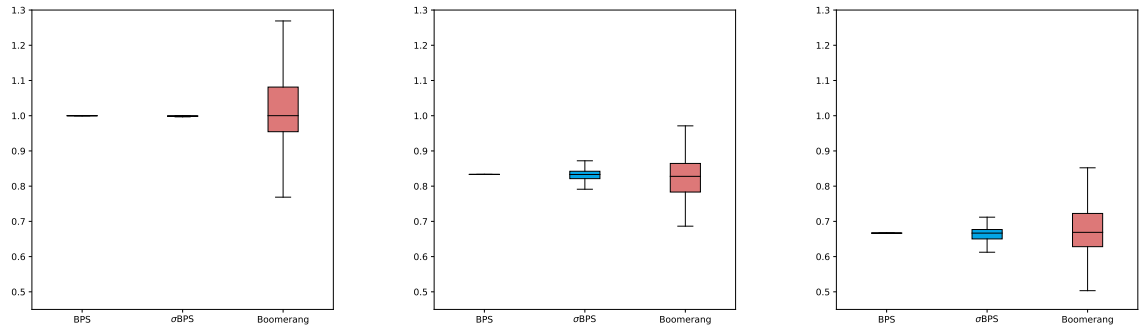
Ethan Goan[1] **Dimitri Perrin**[1] **Kerrie Mengersen**[1] **Clinton Fookes**[1]

[1]Queensland University of Technolgy

A TIGHTNESS OF PROPOSED APPROXIMATE UPPER BOUND

The key contribution within this work is the proposal of a new generic and data-dependent thinning method to approximately sample event times from within PDMP samplers. The quality of this thinning method relies on two key components; the tightness of the envelope and its ability to provide a strict upper bound. We want the envelope used to be able to be as close to the true event rate as possible without reducing below it. This enables maximum efficiency of thinning methods by reducing the likelihood of a proposed event time will be rejected.

Previous works have relied on performing experimentation on simple well-defined models where derivation of a strict and exact upper boundBouchard-Côté et al. [2018], Bierkens et al. [2019, 2020], Wu and Robert [2017]. Derivation for a strict upper bound is infeasible for neural networks, though we can assess the quality of our event thinning method by analysing the acceptance ratio in Equation 8 from the body of the paper. We want this ratio to be as close to one as possible without exceeding it, otherwise, the envelope section used to propose the time is below the true event rate. We assess the distribution of these acceptance ratios for varying values of α from Equation 12 in the body of the paper in Figure 1 and we illustrate the result on predictive performance and computational load in Table 1. We can see that with $\alpha = 1.0$, we see frequent



(a) $\alpha = 1.0$ (b) $\alpha = 1.2$ (c) $\alpha = 1.5$

Figure 1: Distribution of acceptance ratios for event thinning across the different PDMP samplers used within this work for varying levels of α. All models are fit on the MNIST data set as described in Section 5.2.

occurrences of the proposed envelope being below that of the true event rate, though as we increase the value of α, the likelihood of the approximate envelope being a strict upper bound increases. In practice, setting this scaling parameter can be achieved through the use of a small warm-up phase at the start of sampling to find a ratio that satisfies a users willingness to mitigate bias that may be induced due to the violation of the upper bound assumption. To mitigate potential bias, the value of α may be increased at the expense of a small increase in the computational demands of the thinning method as seen

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

Table 1: Summary of predictive performance with and timings as the scaling value of α is increased for the PDMP samplers demonstrated within. All models are fit to the MNIST dataset using the Lenet5 architecture.

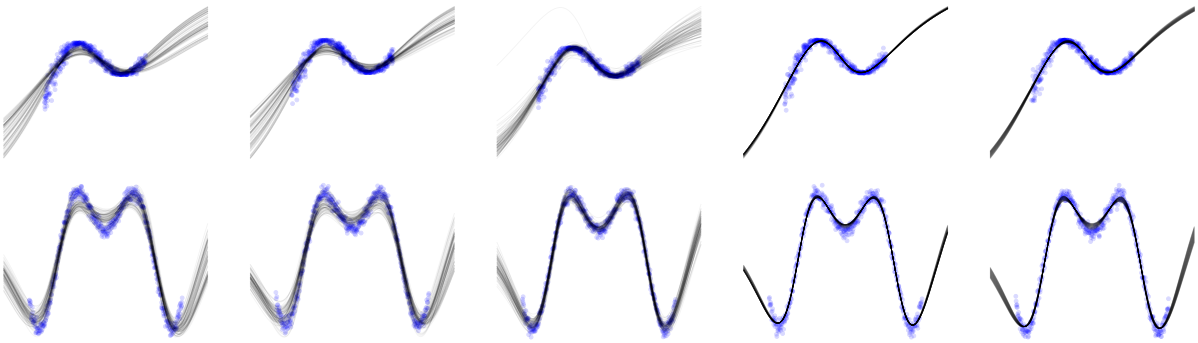
α	**Inference**	**ACC**	**NLL**	**ECC**	**Time**
$\alpha = 1.0$	BPS	0.9914	1.4227	74.752	71
	σBPS	0.9908	0.0375	1.0445	121
	Boomerang	0.9919	0.0230	0.139	77
$\alpha = 1.2$	BPS	0.9906	1.0778	64.4556	75
	σBPS	0.9900	0.2141	16.4637	130
	Boomerang	0.9925	0.0234	0.1736	82
$\alpha = 1.5$	BPS	0.9909	1.0907	64.8491	80
	σBPS	0.991	0.8527	56.0289	143
	Boomerang	0.9922	0.0232	0.1651	86

in Table 1.

B ADDITIONAL REGRESSION AND BINARY CLASSIFICATION EXAMPLES

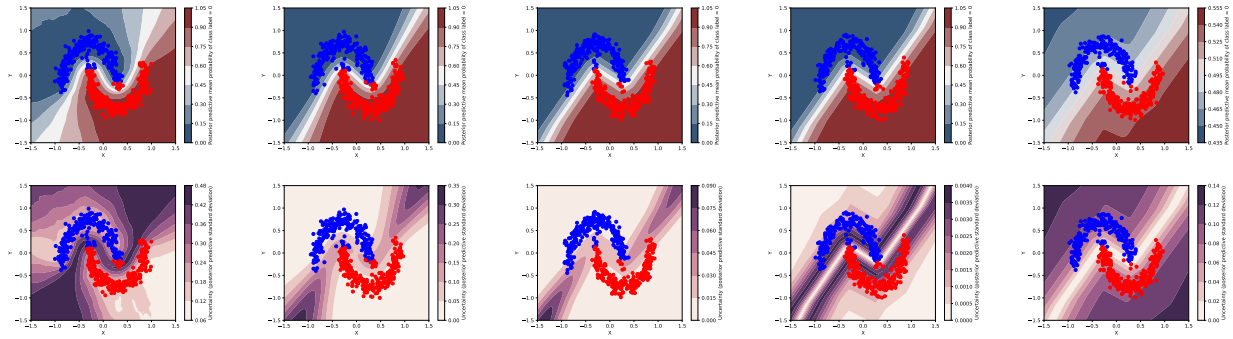
To further validate the predictive performance of PDMP samplers using the proposed event thinning method, we provide additional examples on easy to visualise regression tasks in Figure. 2 and Figure 3 which are compared with Stochastic Gradient Langevin Dynamics (SGLD) with a decreasing learning rate as required, and a constant learning rate with no decay as is typically done in practice (SGLD-ND). For both regression and binary classification models, SGLD experiments are run with a learning rate starting at $1e^{-5}$ and decays to zero linearly. For SGLD-ND, the learning rate of regression models is set to the largest value found that would avoid divergences, resulting in $1e^{-5}$ for regression models and $1e^{3}$ for binary classification.

From these results, we further validate the predictive performance of these samplers and their ability to yield informative uncertainty information for out-of-distribution data when compared to SGLD with a decaying learning rate and a constant learning rate. We find that even with a larger value learning rate used for SGLD-ND that the sampler is unable to explore the posterior sufficiently to provide meaningful uncertainty estimates. This phenomenon has been reported in Brosse et al. [2018], where they identify that even with a larger and constant learning rate, SGLD dynamics converge to that of regular SGD.



(a) BPS (b) σBPS (c) Boomerang (d) SGLD (e) SGLD-ND

Figure 2: Exampled of predictive posteriors for BNN regression models across synthetic data sets. Training samples are shown in blue dots, and draws from the predictive distrubtion shown with black lines.

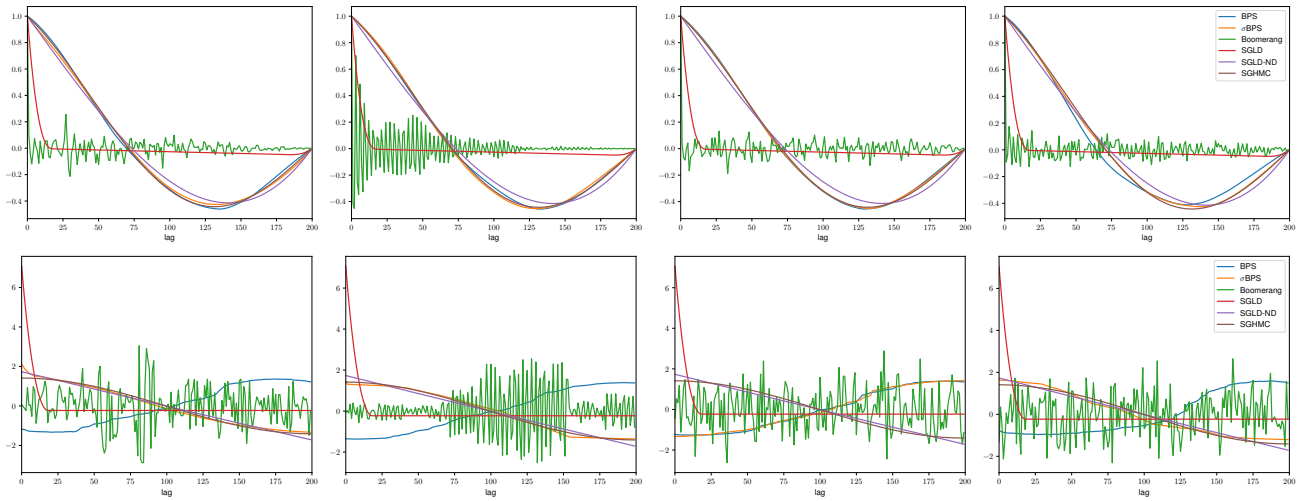


(a) BPS (b) σBPS (c) Boomerang (d) SGLD (e) SGLD-ND

Figure 3: Examples of predictive distributions for synthetic binary classification task. Top row indicates predictive mean and bottom row illustrates variance in predictions. Best viewed on a computer screen in colour.

C MIXING PERFORMANCE

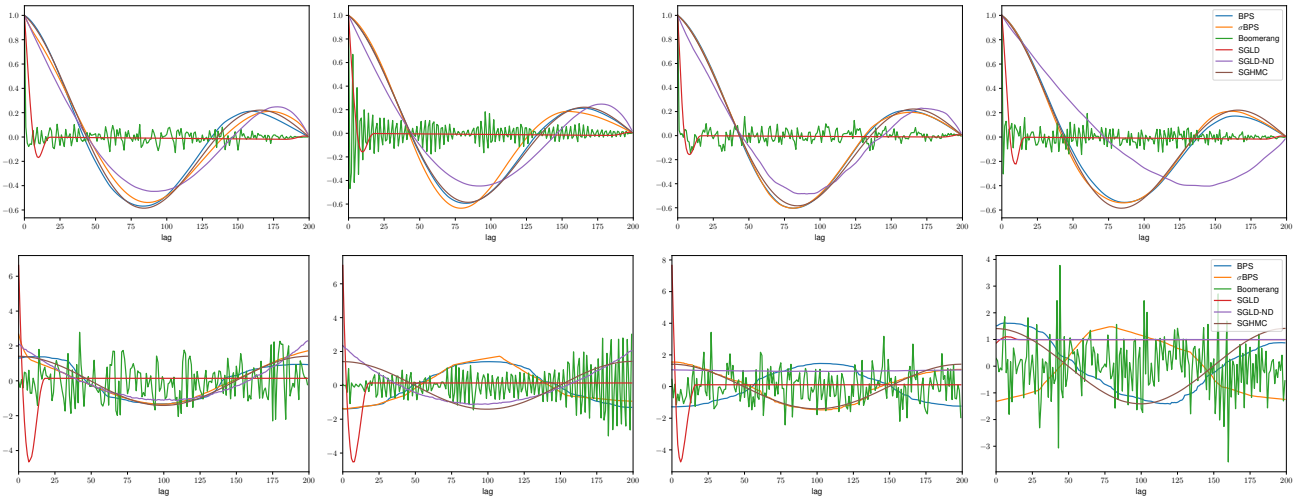
In Section 5.2, experiments to investigate the mixing capabilities of the PDMP samplers were conducted using PCA to reduce the dimensionality of the samples generated from the different samplers for a single network. We extend this analysis here for all models in Figures 4, 5 and 6 for the first, second, and last principal components respectively. From this these figures we can verify that the Boomerang sampler provided the greatest mixing across the different models and datasets, whilst SGLD consistently converges to a single solution. We further investigate this here by comparing just the two samplers for raw parameter traces within different parts of the networks used for the MNIST and SVHN datasets. These results are shown in Figure 7, and confirms the pathology of SGLD quickly converging to a single steady-state solution, whilst the Boomerang sample is able to explore the posterior at all stages in the networks. From this, we can verify that SGLD is



(a) MNIST (b) Fashion-MNIST (c) SVHN (d) CIFAR-10

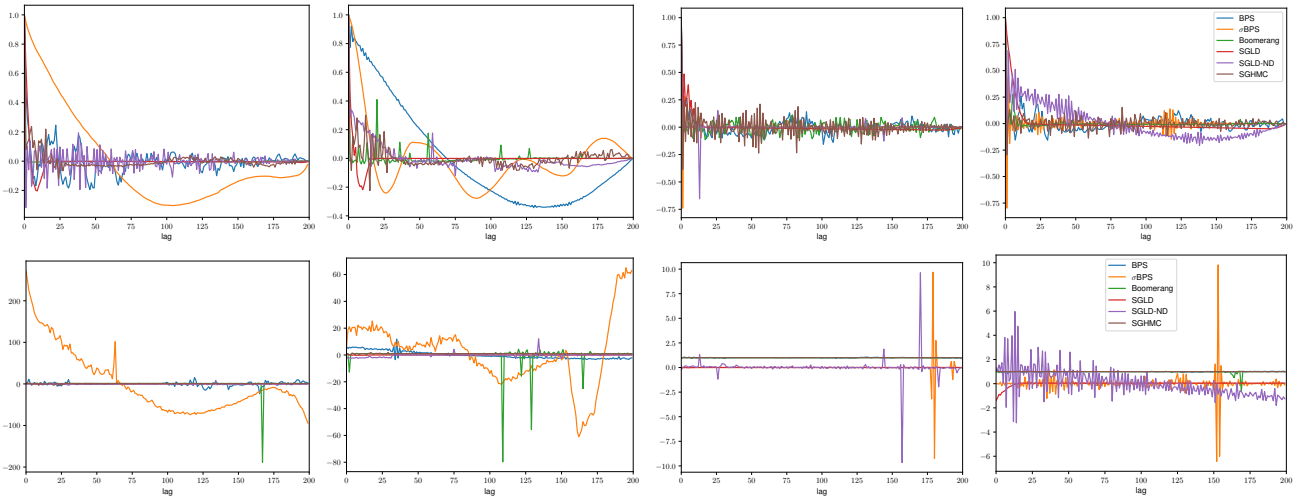
Figure 4: Plots summarising samples from tested samples projected onto first principal component. Top row represents the ACF plot, and the bottom shows the coordinate trace plot for the first principal component. Best viewed on a computer screen.

converging to a steady-state solution, whilst the Boomerang sampler consistently explores the posterior space and provide improved mixing. Given the requirement for SGLD to maintain a small learning rate that approaches 0 to target the posterior Nagapetyan et al. [2017], Brosse et al. [2018], Welling and Teh [2011], these results are expected. The theoretic ability of SGLD to maintain the posterior as its invariant distribution comes at the expense mixing efficiency.



(a) MNIST (b) Fashion-MNIST (c) SVHN (d) CIFAR-10

Figure 5: Plots summarising samples from tested samples projected onto second principal component. Top row represents the ACF plot, and the bottom shows the coordinate trace plot for the first principal component. Best viewed on a computer screen.



(a) MNIST (b) Fashion-MNIST (c) SVHN (d) CIFAR-10

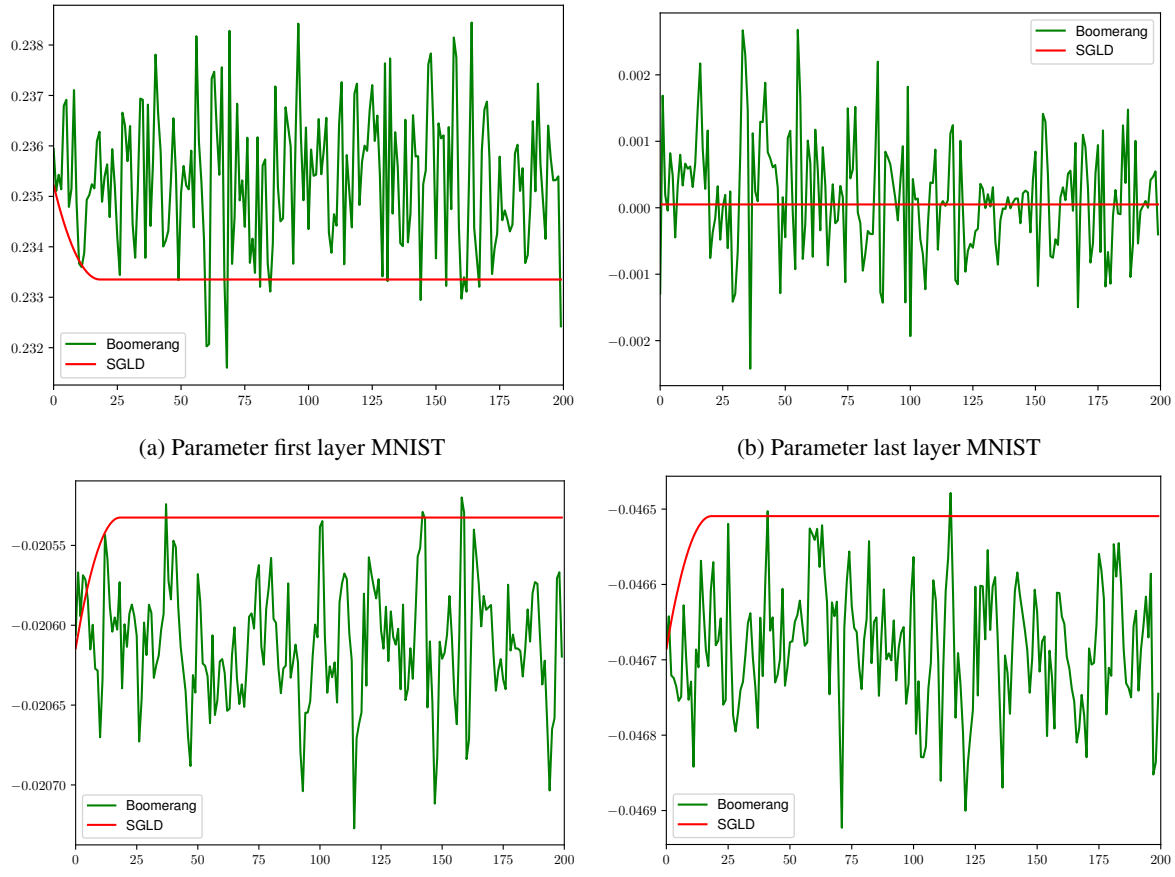
Figure 6: Plots summarising samples from tested samples projected onto last principal component. Top row represents the ACF plot, and the bottom shows the coordinate trace plot for the last principal component. Best viewed on a computer screen.

D SENSITIVITY TO HYPER-PARAMETERS

D.1 SENSITIVITY TO VELOCITY DISTRIBUTION

As noted in Section 6, we discuss the sensitivity of these PDMP Samplers for BNNs with respect to the distribution assigned to the auxiliary velocity variable. Given that the aim of this velocity variable is to guide the dynamics of the system to efficiently explore the parameter space, it needs to be set appropriately. We demonstrate this here through experimentation to highlight how poorly specified velocity distribution can corrupt inference.

Figure 8 illustrates the predictive distribution for poorly specified velocity distributions for the Boomerang sampler, though similar effects are seen amoungst the other PDMP samplers when the variance for the velocity distribution is incorrectly



(a) Parameter first layer MNIST

(b) Parameter last layer MNIST

(c) Parameter first layer SVHN

(d) Parameter last layer SVHN

Figure 7: Trace plots comparing mixing of SGLD and the Boomerang sampler for individual weight parameters within different networks at different locations.

specified. We see that the scale of the velocity component proportionately controls the mixing capabilities of the model. When variance is too low, the sampler is unable to explore beyond the MAP solution, and when too large the predictive performance can suffer. With better approximations to the diagonal of Hessian of the negative log-likelihood, the effects of this may be mitigated for the Boomerang Sampler. We highlight these behaviours of PDMP samplers applied to BNNs to show the limitations and to provide insight into the importance of setting these parameters correctly, and areas of future research.

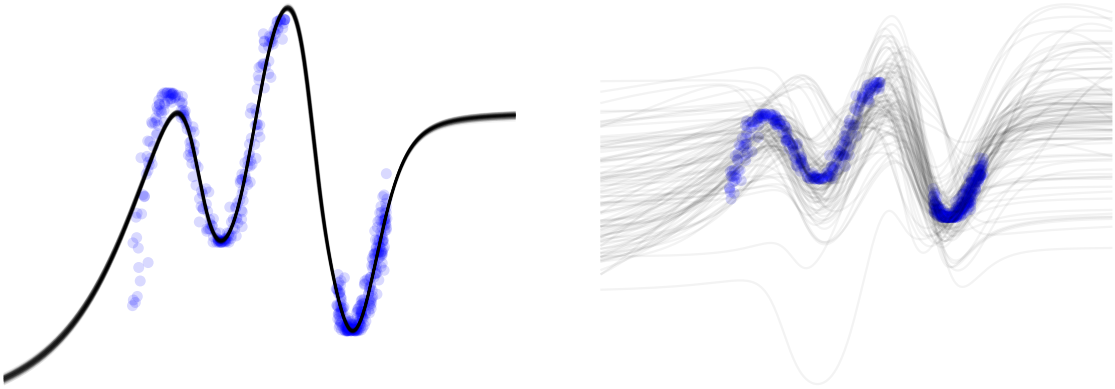
D.2 SENSITIVITY TO REFRESH EVENT RATE

MCMC samplers such as HMC Neal et al. [2011] and NUTS Hoffman et al. [2014] have step size parameters that can be adjusted and tuned for individual models. With a small step size, exploration of the posterior can be limited, and if too large then divergences in the posterior trajectory can be encountered and corrupting inference Betancourt [2017]. The step size parameter is typically tuned during a warm-up phase before sampling is commenced to find an optimal value to maximise exploration and minimise the risk of encountering these divergences.

The PDMP samplers within here do not have an equivalent parameter that can be tuned to guide simulation. The Trajectory of these samplers is defined solely on the transition kernel to update velocity parameters and the event rate that determines when these events occur. We can however yield a similar effect to adjusting the step size of a traditional MCMC model through our choice of event rate for our refreshment process PP(λ_{ref}).

Recall from Section 2.4 that the final event time is given by,

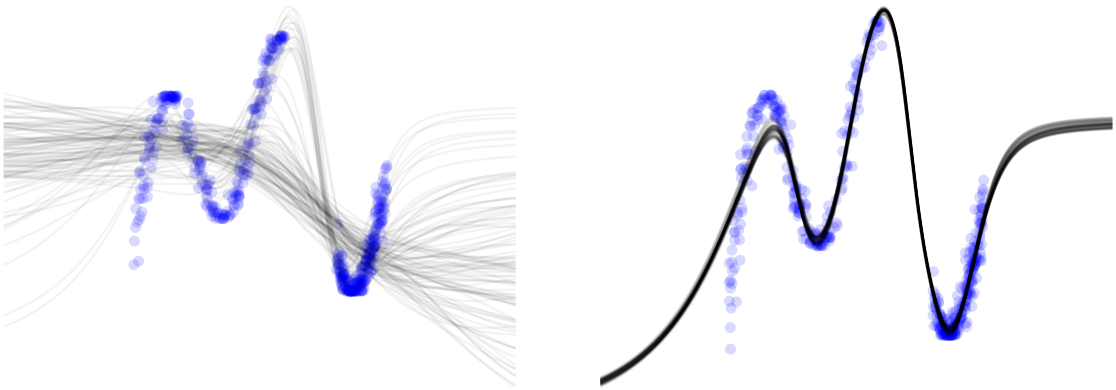
$$\tau_{event} = \min(\tau, \tau_{ref}) \tag{1}$$



(a) Small velocity distribution $\gamma = 1$ (b) Large velocity distribution $\gamma = 1000$

Figure 8: Effect of scale in velocity reference measure for PDMP samplers applied to BNNs.

Where $\tau \sim \text{PP}(\lambda(\omega(t), \mathbf{v}(t)))$, and $\tau_{ref} \sim \text{PP}(\lambda_{ref})$. Setting the value for τ_{ref} can implicitly control the level of exploration within our samplers. For large λ_{ref}, we will encounter smaller proposed refresh times and thus will refresh more frequently. Similarly, for larger λ_{ref}, our sampled refresh times will be larger, and τ_{event} will equal τ more frequently, and further exploration of the posterior space with these dynamics will be possible. We illustrate this in Figure 9, where we show the effects for large and smaller values of λ_{ref}.



(a) $\lambda_{ref} = 0.01$ (b) $\lambda_{ref} = 10.0$

Figure 9: Effect of λ_{ref} on PDMP models applied to BNNs. Shown here is the predictive distribution found with the BPS using the proposed event rate thinning method.

We can see that the refresh rate can have a considerable impact on the inference quality of our model. With λ_{ref} too large, our exploration is limited and we perform excessive refreshments instead of accepting those provided by the PDMP kernel. When is too small, we can accept larger event times as specified by the PDMP sampler and can diverge away from meaningful inferences.

E SUMMARY OF MODELS USED

E.1 REGRESSION AND BINARY CLASSIFICATION MODELS FOR SYNTHETIC DATA

Regression models used within this work consist of fully-connected networks with two hidden layers, each with 25 and 10 units respectively. Tanh non-linear activations are applied after each hidden layer, and a Normal likelihood with a variance

of $\sigma^2 = 0.01$ is used. MAP estimates for these networks are found with 10,000 iterations using the Adam optimiser Kingma and Ba [2015], with each sampler initialised from the same MAP estimate.

For binary classification models, the networks consist of a fully-connected network with three hidden layers, each with 100 units. ReLU non-linear activations are applied within the network, and a Bernoulli likelihood is used. Similar to the regression tasks, MAP estimate is found with Adam.

E.2 ADDITONAL UCI-DATASET RESULTS

We provide here additional results on datasets from the UCI repositoryNewman et al. [1998]. For each dataset, a simple MLP network with a single hidden layer with 50 hidden units is used, along with a Tanh activation. MAP estimates are found similar to , followed by 2,000 samples generated by each method. Each experiment is run 5 times with means results and standard deviations reported. We further include ESS as measured on from the smallest principle component of samples. Results from these experiments reflect that seen in Section 5.1.1; where SGLD is able to provide an almost negligible improvement in MSE and NLL, though is unable to provide efficient posterior exploration. The Boomerang sampler is able to consistently outperform other sampling methods, with other samplers only able to match sample efficiency for the smallest principal components where exploration is smallest.

Table 2: Results on UCI-Naval Dataset

Inference	NLL	MSE	ESS-First	ESS-Last
BPS	1681.66 ± 0.64	$\mathbf{0.01 \pm 0.00}$	2.73 ± 0.02	$\mathbf{2000.00 \pm 0.00}$
σBPS	1685.43 ± 4.97	$\mathbf{0.01 \pm 0.00}$	2.71 ± 0.02	$\mathbf{2000.00 \pm 0.00}$
Boomerang	1680.55 ± 0.08	$\mathbf{0.01 \pm 0.00}$	$\mathbf{1777.57 \pm 156.57}$	1804.77 ± 219.12
SGLD	$\mathbf{1680.47 \pm 0.00}$	$\mathbf{0.01 \pm 0.00}$	2.88 ± 0.03	156.24 ± 72.36
SGHMC	1689.88 ± 7.88	$\mathbf{0.01 \pm 0.00}$	2.72 ± 0.01	$\mathbf{2000.00 \pm 0.00}$

Table 3: Results on UCI Energy Dataset

Inference	NLL	MSE	ESS-First	ESS-Last
BPS	74.09 ± 0.10	$\mathbf{2.96 \pm 0.04}$	2.70 ± 0.01	1795.50 ± 49.44
σBPS	74.09 ± 0.08	2.97 ± 0.04	2.69 ± 0.03	1527.35 ± 115.94
Boomerang	74.20 ± 0.03	3.01 ± 0.01	$\mathbf{1981.29 \pm 25.87}$	1880.46 ± 178.71
SGLD	$\mathbf{74.06 \pm 0.00}$	$\mathbf{2.96 \pm 0.00}$	2.87 ± 0.00	272.34 ± 247.11
SGHMC	74.36 ± 0.18	3.08 ± 0.08	2.72 ± 0.02	$\mathbf{1992.02 \pm 17.84}$

Table 4: Results on UCI Yacht Dataset

Inference	NLL	MSE	ESS-First	ESS-Last
BPS	32.34 ± 0.03	7.55 ± 0.03	2.72 ± 0.02	1467.26 ± 11.96
σBPS	32.33 ± 0.08	7.54 ± 0.07	2.72 ± 0.03	1291.29 ± 52.81
Boomerang	32.41 ± 0.11	7.62 ± 0.11	$\mathbf{2000.00 \pm 0.00}$	$\mathbf{1945.22 \pm 61.12}$
SGLD	32.32 ± 0.00	7.53 ± 0.00	2.87 ± 0.00	3.29 ± 0.58
SGHMC	$\mathbf{32.27 \pm 0.17}$	$\mathbf{7.48 \pm 0.17}$	2.73 ± 0.02	1652.04 ± 30.10

E.3 CONVOLUTIONAL MODELS

For the σBPS, an initial warm-up stage is again required, which is identical to that in Section 5.1. For MNIST and Fashion-MNIST, a batch size of 1024 is used, whilst a batch size of 512 is used for the remaining models. MAP estimates for MNIST and Fashion-MNIST datasets were found with the Adam optimiser Kingma and Ba [2015] for 10,000 iterations. SVHN and

Table 5: Results on UCI Concrete Dataset

Inference	NLL	MSE	ESS-First	ESS-Last
BPS	111.51 ± 0.12	9.55 ± 0.03	2.72 ± 0.03	1777.00 ± 112.52
σBPS	111.60 ± 0.27	9.58 ± 0.08	2.72 ± 0.02	1503.21 ± 36.08
Boomerang	111.95 ± 0.35	9.68 ± 0.10	$\mathbf{1975.76 \pm 54.21}$	$\mathbf{1982.03 \pm 24.85}$
SGLD	$\mathbf{111.47 \pm 0.00}$	$\mathbf{9.54 \pm 0.00}$	2.87 ± 0.00	88.94 ± 118.56
SGHMC	112.18 ± 0.61	9.74 ± 0.17	2.72 ± 0.02	1906.15 ± 128.18

CIFAR-10 used SGD with momentum of 0.1 and 0.9 respectively for 25,000 iterations, where for CIFAR-100, required 128,000 iterations and a momentum of 0.2.
With the potential sensitivities to both refreshment rates and choice of velocity distribution $\Phi(\mathbf{v})$ identified in D, we deem it important to report the values used for fitting each model. We report these values in Table 6 alongside full predictive performance measurements and sample efficiency metrics. Within Table 6, we represent the choice of velocity distribution with the γ parameter. For Bouncy Particle Sampler (BPS) and σBPS, γ describes the standard deviation of the velocity distribution chosen such that,

$$\Phi(\mathbf{v}) = \mathcal{N}(0, \gamma^2). \tag{2}$$

For the Boomerang sampler, γ represents the scaling factor as found in Equation 5 from the body of the paper. Included in these results is the Effective Sample Size (ESS) when projecting the samples onto the first, second, and last principal components. We see from these results that the Boomerang Sampler can generate independent samples across all components, whilst other samplers are only able to offer this independence as the amount of information (or variance) in the projection of these samples decreases.

F HOW WELL ARE WE REALLY EXPLORING THE POSTERIOR?

In Radford Neals influential thesis Neal [2012], he states that "Bayesian neural network users may have difficulty claiming with a straight face that their models and priors are selected because they are just what is needed to capture their prior beliefs about the problem" [1]. In a similar vein, we would state that any Bayesian neural network user would have a difficult time honestly saying their inference strategy has sufficiently explored the posterior, including the work proposed here. Previous research has investigated gold-standard MCMC methods for larger networks Izmailov et al. [2021], though were unable to obtain a sufficient number of samples to maintain confidence in the levels of exploration. Although the metrics in the previous section may show sufficient results for a machine learning application, from a statistical perspective we need to further investigate the quality of our inference to justify whether we have satisfied our goal of sampling from the posterior distribution.

Previous papers for PDMP methods for MCMC have shown favourable performance in terms of mixing and sampling efficiency Bouchard-Côté et al. [2018], Bierkens et al. [2019], Wu and Robert [2017], Bierkens et al. [2020] and has similarly outperformed methods such as SGLD. Most studies have been restricted to well-defined models; where prior information can be suitably provided and sufficient prior studies with gold standard methods such as HMC and NUTS have confirmed the general geometry of the posteriors in question and the existence of a central limit theorem. Inference in BNNs is challenged by a posterior with strong multi-modality, making exploration of any sampler more difficult. This is further challenged by the comparatively large dimensional space over which we need to explore. The favourable Gaussian Process and functional properties seen by networks with infinite width Neal [2012] encourage the use of large models, whilst also narrowing the typical set in which we wish to explore Betancourt [2017].

Another limitation is the computational complexity added with sampling-based schemes. This complexity not only includes the cost of sampling, but the increase in memory consumption. The popular ResNet-50 model contains more than 23 million parameters. To perform inference on ImageNet with a ResNet50 model using a mini-batch size of 100 samples, more than 10 thousand samples would be needed to iterate over the entire data set of over one million images. With single-point precision,

[1] Although much important work has been conducted to establish suitable priors and model design Hafner et al. [2018], Sun et al. [2019], Vladimirova et al. [2019], this statement largely remains true today.

[1] The commonly used variant of ImageNet is from the 2012 Large Scale Visual Recognition Challenge, which contains 1,281,167 samples Russakovsky et al. [2015]

Table 6: Summary of hyperparameters used for samplers within this work.

Dataset	Inference	λ_{ref}	γ	Time
MNIST	SGD	-	-	74
	SGLD	-	-	87
	SGLD-ND	-	-	87
	BPS	1.0	0.001	145
	σBPS	1.0	0.25	197
	Boomerang	0.1	0.01	151
Fashion-MNIST	SGD	-	-	74
	SGLD	-	-	87
	SGLD-ND	-	-	87
	BPS	1.0	0.001	144
	σBPS	0.1	0.001	192
	Boomerang	0.1	0.01	156
SVHN	SGD	-	-	3465
	SGLD	-	-	3653
	SGLD-ND	-	-	3653
	BPS	1.0	0.001	4125
	σBPS	0.1	0.0001	4535
	Boomerang	1.0	0.0001	4375
CIFAR 10	SGD	-	-	4905
	SGLD	-	-	5075
	SGLD-ND	-	-	5074
	BPS	1.0	0.001	5614
	σBPS	0.1	0.0001	6053
	Boomerang	0.1	0.001	5868
CIFAR 100	SGD	-	-	9811
	SGLD	-	-	9985
	SGLD-ND	-	-	9985
	BPS	1.0	0.001	10478
	σBPS	0.1	0.001	10808
	Boomerang	0.1	0.001	10783

these samples for a single complete iteration of the dataset would require more than 9.2GB of memory. These constraints currently limit the applicability of such methods, as evaluating predictive posteriors will require a large number of samples and many operations to read sampled values from non-volatile storage.

These limitations offer insights into areas of future research relating to sampling schemes for BNNs. The geometry of the joint posterior distribution could be improved by investigating non-local methods for preconditioning the gradients, similar to that done in Riemannian HMC Girolami and Calderhead [2011]. As seen in this work through the efficacy of the Boomerang sampler, exploitation of this geometry can considerably improve exploration of the posterior space. Finally and perhaps most importantly, bespoke design of model architecture that respects the data and includes priors that appropriately reflect domain expertise could yield posteriors that are more easily traversed and explored.

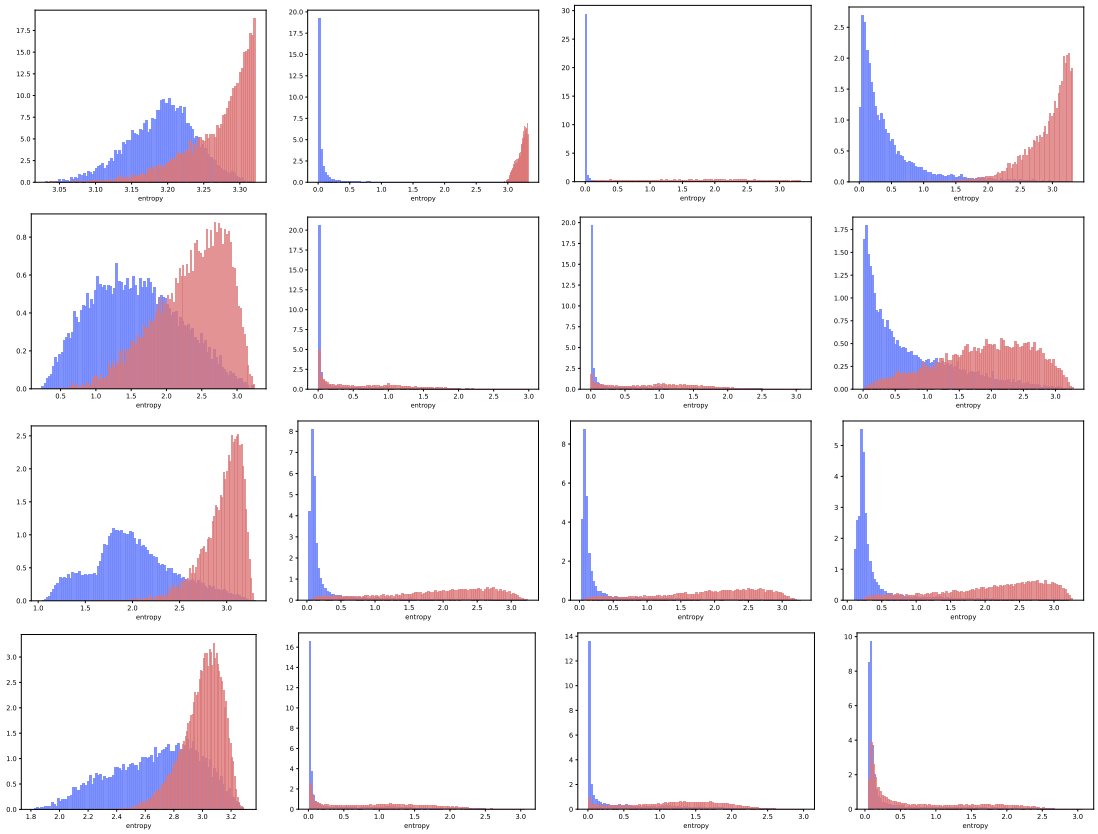
G IN AND OUT OF DISTRIBUTION DATA

We investigate here the performance of the different sampling methods for in and out-of-distribution (OOD) data in terms of predictive classification entropy. We have demonstrated that PDMP sampling methods present meaningful epistemic uncertainty in predictions. It is important to identify uncertainty in the final predictions that are made. Within this work, predictions are made by taking the argmax of the mean for the predictive posterior,

$$\mathbf{t} = \underset{y^* \in \mathcal{Y}}{\mathrm{argmax}} \ \mathbb{E}_{y^*} \left[p(y^*|x^*, \mathcal{D}) \right] \tag{3}$$

Where $p(y^*|x^*, \mathcal{D})$ is our predictive posterior. Entropy within this categorical probability vector given by this expectation can be viewed as an approximate measure for aleatoric uncertainty within our model Smith and Gal [2018] to accompany

the epistemic uncertainty given by our Bayesian models. To assess this, we compute the entropy of the expectation within Equation 3 for in-distribution data and OOD data. It is desirable to have a lower entropy for in-distribution data indicating lower predictive aleatoric uncertainty, and a larger entropy for OOD data to represent an increase in uncertainty. Figure 10 illustrates this for the models used within this work. From Figure 10, we see the BPS provides increased entropy for



(a) BPS (b) σBPS (c) Boomerang (d) SGLD

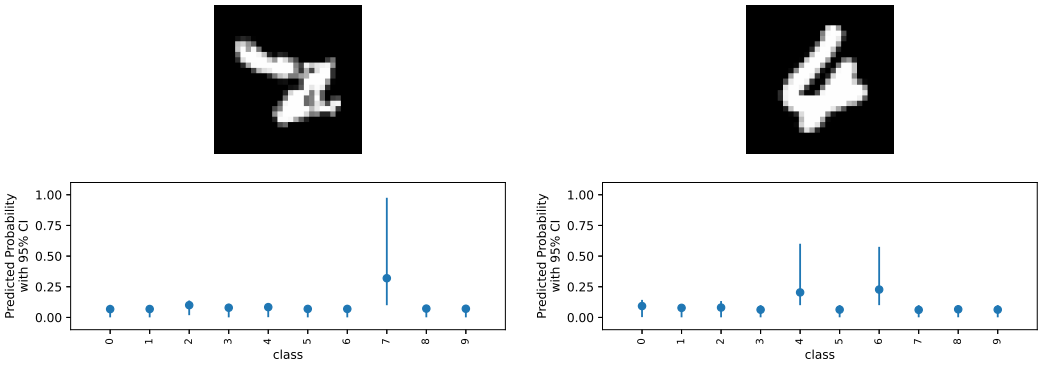
Figure 10: Entropy within the final predictive categorical vector obtained from the tested sampling methods for the different datasets used. Blue histograms indicate in-data-distribution entropy and red for OOD data. Each column represents the predictive entropy for the corresponding labelled sampler and each row for a different dataset. From top to bottom, each row is for models fit on the MNIST, Fashion MNIST, SVHN, and CIFAR-10 data respectively. MNIST and Fashion MNIST datasets are used to model in and OOD datasets for the applicable models, and similarly SVHN and CIFAR-10 to model in and OOD for respective models.

OOD data though is unable to provide a small entropy for in-distribution data. Results from the σBPS, Boomerang, and SGLD samplers all provide similar trends and provide increased uncertainty for OOD data, however for in-distribution data, the Boomerang sampler provides a reduction in epistemic uncertainty. A low predictive entropy for in-distribution data could indicate overconfidence and should not be used in isolation to evaluate calibration. Combining these results with the improved ECE calibration scores obtained with predictions from the Boomerang sampler indicate favourable predictive performance; predictions from these results are well-calibrated for in-distribution and showing appropriately reduced aleatoric uncertainty, whilst providing comparable or improved predictive uncertainty for OOD data.

H EXAMPLES OF DIFFICULT TO CLASSIFY SAMPLES

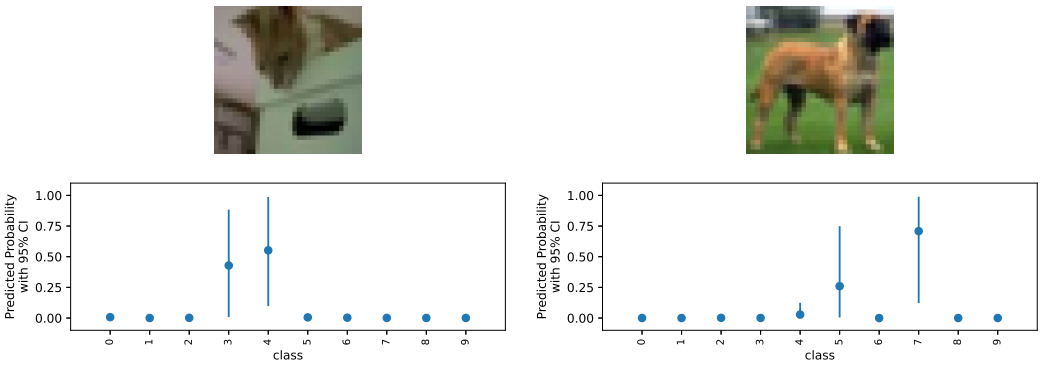
Given the increasing desire to apply deep learning models in practice, the ability for them to reliably communicate uncertainty information is crucial. We expect our models to encounter difficult-to-understand scenarios. We need to be able to identify when these challenging scenarios occur and to incorporate the uncertainty encountered into final decisions. Figure 11 illustrates examples of misclassified samples from the datasets evaluated within this work, and illustrates the predictive probabilities of these models. We see that all PDMP samplers provide meaningful uncertainty information for

difficult-to-classify instances within each data set.



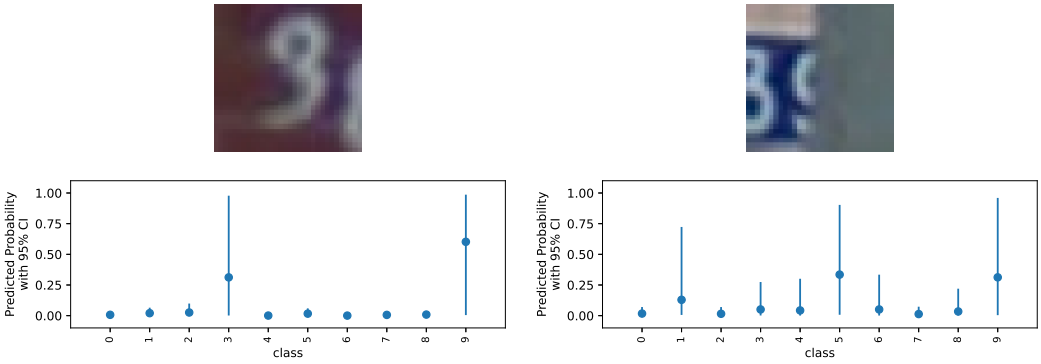
(a) ‘Two‘ from MNIST misclassified as ‘Seven‘

(b) ‘Four‘ from MNIST missclassified as ‘Six‘



(c) ‘Cat‘ from CIFAR-10 misclassified as ‘Deer‘

(d) ‘Dog‘ from CIFAR-10 misclassified as ‘Horse‘



(e) ‘Three‘ from SVHN missclassified as ‘Nine‘

(f) ‘Nine‘ from SVHN missclassified as ‘five‘

Figure 11: Examples of difficult-to-classify images from the different image data sets used. Below each image is the predictive mean for each class represented by the dot, and error bars to represent the 95% credible intervals. MNIST results fit with BPS, CIFAR-10 with σBPS, and SVHN with Boomerang sampler using the proposed event thinning method. Best viewed on a computer screen.

References

Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Joris Bierkens, Paul Fearnhead, Gareth Roberts, et al. The zig-zag process and super-efficient sampling for bayesian analysis of big data. *The Annals of Statistics*, 47(3):1288–1320, 2019.

Joris Bierkens, Sebastiano Grazzi, Kengo Kamatani, and Gareth Roberts. The boomerang sampler. *arXiv preprint arXiv:2006.13777*, 2020.

Alexandre Bouchard-Côté, Sebastian J Vollmer, and Arnaud Doucet. The bouncy particle sampler: A nonreversible rejection-free markov chain monte carlo method. *Journal of the American Statistical Association*, 113(522):855–867, 2018.

Nicolas Brosse, Alain Durmus, and Eric Moulines. The promises and pitfalls of stochastic gradient langevin dynamics. In *Advances in Neural Information Processing Systems*, pages 8268–8278, 2018.

Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Danijar Hafner, Dustin Tran, Timothy Lillicrap, Alex Irpan, and James Davidson. Reliable uncertainty estimates in deep neural networks using noise contrastive priors. *arXiv preprint arXiv:1807.09289*, 2018.

Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Wilson. What are bayesian neural network posteriors really like? *arXiv preprint arXiv:2104.14421*, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6980`.

Tigran Nagapetyan, Andrew B Duncan, Leonard Hasenclever, Sebastian J Vollmer, Lukasz Szpruch, and Konstantinos Zygalakis. The true cost of stochastic gradient langevin dynamics. *arXiv preprint arXiv:1706.02692*, 2017.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz. Uci repository of machine learning databases, 1998. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. In *Uncertainty in AI*, 2018.

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational bayesian neural networks. *arXiv preprint arXiv:1903.05779*, 2019.

Mariia Vladimirova, Jakob Verbeek, Pablo Mesejo, and Julyan Arbel. Understanding priors in bayesian neural networks at the unit level. In *International Conference on Machine Learning*, pages 6458–6467. PMLR, 2019.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

Changye Wu and Christian P Robert. Generalized bouncy particle sampler. *arXiv preprint arXiv:1706.04781*, art. arXiv:1706.04781, Jun 2017.