

Reproducing the paper: *Stochastic Gradient Hamiltonian Monte Carlo* by Tianqi Chen, Emily B. Fox and Carlos Guestrin

Sam Adam-Day, Alexander Goodall, Theo Lewy and Fanqi Xu

Abstract

We reproduce the experiments contained in ‘Stochastic Gradient Hamiltonian Monte Carlo’ [CFG14b] by Chen, Fox and Guestrin.

Fixme: Give more details in abstract

List of Corrections

Give more details in abstract	1
Give this a citation	4

1 Introduction

Hamiltonian Monte Carlo (HMC) methods already provide a way to sample from a posterior distribution. These methods use all data available to them to produce a potential energy function:

$$U(\theta) = - \sum_{x \in \mathcal{D}} \log p(x|\theta) - \log p(\theta) \propto -\log p(\theta|\mathcal{D})$$

which, along with $\nabla U(\theta)$, is sufficient to sample from the posterior (as will be discussed in the background section). Computing $U(\theta)$ and $\nabla U(\theta)$ can be computationally expensive for large datasets, and so research into Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) methods began with [CFG14b], which use subsets of the data to produce noisy estimates of $U(\theta)$ and $\nabla U(\theta)$. We decided to investigate the paper that introduced SGHMC. It starts with a description of HMC, and then introduces a Naive SGHMC algorithm. It then goes on to introduce SGHMC (the main algorithm), and then it runs a number of experiments using these algorithms.

Our motivation to choose this paper stemmed from the fact that the SGHMC algorithm is relatively simple, and so we would have more time to investigate other datasets and ideas. It is also computationally fast, allowing us to run experiments on our home computers. We reproduced the following experiments from [CFG14b]:

- Sampling θ from the posterior with $U(\theta) = -2\theta^2 + \theta^4$ using HMC, Naive SGHMC and SGHMC
- Sampling (θ, r) generated from $U(\theta) = \frac{1}{2}\theta^2$ with HMC, and from $U(\theta) = \frac{1}{2}\theta^2 + \mathcal{N}(0, 4)$ as a proxy for SGHMC.

- Classifying the MNIST dataset [Den12] using SGHMC as well as with Stochastic Gradient Descent, Stochastic Gradient Descent with momentum, and Stochastic Gradient Langevin Dynamics.

We also considered some new ideas:

- We extended the ‘No U-Turn Sampler’ (NUTS) from [HG11] to work with SGHMC to produce our novel algorithm SGNUTS.
- We ran experiments on a new dataset of FashionMNIST [XRV17].
- We briefly introduced some Convolutional Neural Networks (CNNs) to see how accurate SGHMC was at classifying CIFAR10 [Kri09].
- We attempted to evaluate the noisiness of the data (B in the literature and in what follows) and used this to increase the algorithm’s efficiency.

The repository for our code is found at <https://github.com/sacktock/SGHMC>.

2 Background

2.1 HMC

Hamiltonian Monte Carlo ([Dua+87; Nea11]) is a Markov Chain Monte Carlo (MCMC) sampling algorithm. Given a target probability distribution — in our case the posterior distribution of a set of variables θ given independent observations $x \in \mathcal{D}$ — it produces samples by carrying out a random walk over the parameter space using Hamiltonian dynamics.

We begin with the prior distribution $p(\theta)$ and likelihood $p(x | \theta)$. Using these we define the *potential energy* function U :

$$U(\theta) := - \sum_{x \in \mathcal{D}} \log p(x | \theta) - \log p(\theta)$$

Note that, using Bayes’ rule, we have that the posterior $p(\theta | \mathcal{D}) \propto \exp(-U)$. Hamiltonian dynamics introduces an auxiliary set of momentum variables r . These dynamics have a physical interpretation in which an object moves about a landscape determined by U . We let this object have *mass matrix* M , which is typically set to the identity matrix. Then $U(\theta)$ represents the potential energy of the object, and its kinetic energy is given by $\frac{1}{2}r^T M^{-1} r$. The total energy of the system is a quantity known as the *Hamiltonian function*:

$$H(\theta, r) = U(\theta) + \frac{1}{2}r^T M^{-1} r$$

The development of the system is governed by the following equations.

$$\begin{aligned} d\theta &= M^{-1}r dt \\ dr &= -\nabla U(\theta) dt \end{aligned}$$

To simulate these continuous dynamics in practice, we must use a discretised version of these equations. To correct for the inaccuracies introduced by doing so, it is necessary to make a *Metropolis-Hastings correction step*. A simple algorithm is given in Algorithm 1.

In practice, the dataset \mathcal{D} may be large, and so running Algorithm 1 may be computationally expensive, as the complexity of calculating ∇U scales with $|\mathcal{D}|$. One idea to combat this is to simulate the Hamiltonian system using only a subset of the data at a time, in analogy with stochastic gradient descent. This method is known as Stochastic Gradient Hamiltonian Monte Carlo (SGHMC), however before explaining the SGHMC algorithm we begin with Naive SGHMC.

Algorithm 1 A simple HMC algorithm

```

for  $t = 1, 2, \dots$  do
     $r \sim \mathcal{N}(0, 1)$  ▷ Resample momentum
     $(\theta_0, r_0) = (\theta, r)$ 
    for  $i = 1$  to  $m$  do
         $\theta \leftarrow \theta + \epsilon M^{-1}r$ 
         $r \leftarrow r - \epsilon \nabla U(\theta)$ 
    end for
     $u \sim \text{Uniform}[0, 1]$ 
     $\rho = \exp(H(\theta, r) - H(\theta_0, r_0))$  ▷ Acceptance probability
    if  $u < \min(1, \rho)$  then ▷ Only accept new state with probability  $\rho$ 
         $\theta = \theta_0$ 
    end if
end for

```

2.2 Naïve SGHMC

To understand the Naïve SGHMC method, consider sampling a minibatch $\tilde{\mathcal{D}} \subset \mathcal{D}$ uniformly at random. We estimate the gradient $\nabla U(\theta)$ using this minibatch as follows:

$$\nabla \tilde{U}(\theta) = -\frac{|\mathcal{D}|}{|\tilde{\mathcal{D}}|} \sum_{x \in \tilde{\mathcal{D}}} \nabla \log p(x | \theta) - \nabla \log p(\theta)$$

Appealing to the Central Limit Theorem, we imagine that the noisy estimate $\nabla \tilde{U}(\theta)$ is normally distributed about $\nabla U(\theta)$, with some covariance matrix $V(\theta)$. The naïve adaptation of HMC to this stochastic scenario simply replaces $\nabla U(\theta)$ with $\nabla \tilde{U}(\theta)$ in Algorithm 1. The corresponding discrete system is then the ϵ -discretisation of the following dynamics:

$$\begin{aligned} d\theta &= M^{-1}r dt \\ dr &= -\nabla U(\theta) dt + \mathcal{N}(0, 2B(\theta)dt) \end{aligned}$$

where $B(\theta) = \frac{1}{2}\epsilon V(\theta)$.

Unfortunately, such a dynamical system can diverge quite rapidly from the true posterior distribution [Nea11], which necessitates frequent Metropolis-Hastings steps. Such steps are costly since calculating the acceptance probability requires the use of the whole dataset. The method ‘Stochastic Gradient Hamiltonian Monte Carlo’ (SGHMC) proposed in [CFG14b] addresses this shortcoming. The idea is to incorporate friction into the dynamical system, which works to counteract the noise introduced by selecting a subset of the data.

2.3 SGHMC

The SGHMC method adds a ‘friction term’ $BM^{-1}r$ to the momentum update. Since we are unlikely to know the noise model B in practice, we instead take an estimate \hat{B} of B , together with a user-specified friction term $C \succeq \hat{B}$ and simulate the following dynamics.

$$\begin{aligned} d\theta &= M^{-1}r dt \\ dr &= -\nabla U(\theta) dt - CM^{-1}r dt + \mathcal{N}(0, 2(C - \hat{B}(\theta))dt) + \mathcal{N}(0, 2B(\theta)dt) \end{aligned}$$

The algorithm is given in Algorithm 2. In the case where $\hat{B} = B$, these dynamics accurately traverse from the posterior distribution. In practice, we must rely on

an inaccurate estimate \widehat{B} . The simplest choice is $\widehat{B} = 0$. A better but more costly estimate is $\widehat{B}(\theta) = \frac{1}{2}\epsilon\widehat{V}(\theta)$, where \widehat{V} is the observed (empirical Fisher) information [AKW12]. The friction term C can then be taken as a hyperparameter, and set so as to counteract the inaccuracies of the estimate \widehat{B} .

Algorithm 2 The SGHMC algorithm

```

for  $t = 1, 2, \dots$  do
     $r \sim \mathcal{N}(0, 1)$  ▷ Resample momentum
    for  $i = 1$  to  $m$  do
         $\theta \leftarrow \theta + \epsilon M^{-1}r$ 
         $r \leftarrow r - \epsilon \nabla \tilde{U}(\theta) - \epsilon CM^{-1}r + \mathcal{N}(0, 2(C - \widehat{B}(\theta))\epsilon)$ 
    end for
end for

```

3 Implementation Details

We implemented the following algorithms from scratch: HMC, SGHMC, SGLD (stochastic gradient Langevin dynamics [WT11]), SGD (stochastic gradient descent), SGD with Nesterov momentum and SGNUTS — a novel extension to SGHMC that uses ideas from the popular No U-Turn Sampler [HG11]. All of our implementations subclass Pyro’s `MCMCKernel` and are designed to be used directly with Pyro [Bin+19] — a universal probabilistic programming language (PPL) written in Python. In Pyro the user specifies a model which is a probabilistic program (PP) that describes a posterior distribution $p(\theta | \mathcal{D})$ that we want to sample from; θ corresponds to the sampled parameters or latent parameters of the model and \mathcal{D} corresponds to the observed parameters of the model.

Pyro already comes with the following MCMC samplers: HMC, MH and NUTS. So while the algorithms we implemented are not novel they are in fact innovative since Pyro doesn’t come with any of them already built. The main reason for choosing to implement our algorithms on top of Pyro is because Pyro comes with a method `initialize_model` that given a Pyro PP or model P transforms it into a potential function U . Once we have U we can pass the latent and observed parameters (θ, \mathcal{D}) to U , which computes the negative log joint $-\log p(\theta, \mathcal{D})$. Bayes’ rule tells us that $p(\theta | \mathcal{D}) \propto p(\theta, \mathcal{D})$ which is typically all we need for MCMC samplers and even simpler ones such as [Importance and Rejection samplers](#). Pyro also comes with the method `potential_grad`, which given (θ, \mathcal{D}) computes the gradient of U with respect to the parameters θ . The result is that we can specify any arbitrary PP and apply our algorithms to them — letting Pyro handle the transformation from PP to potential function and the gradient computations.

Fixme: Give this a citation

In traditional MCMC samplers the observed dataset \mathcal{D} is constant and so once a Pyro PP has been transformed into a potential function $U(\theta) = -\log p(\theta, \mathcal{D})$ we need not change it. Unfortunately this is less straight forward for stochastic gradient samplers such as SGHMC and SGLD since we subsample the full dataset \mathcal{D} by sampling minibatches $\tilde{\mathcal{D}}$, where $\tilde{\mathcal{D}} \subset \mathcal{D}$. In both SGHMC and SGLD we require that the potential function has the form:

$$\tilde{U}(\theta) = -\frac{|D|}{|\tilde{\mathcal{D}}|} \log p(\tilde{\mathcal{D}} | \theta) - \log p(\theta)$$

Unfortunately when we subsample the dataset and call `initialize_model`

Pyro doesn't explicitly supply us with the likelihood term $p(\tilde{\mathcal{D}} | \theta)$ - it only gives us the negative log joint $-\log p(\theta, \tilde{\mathcal{D}})$, so we had to modify Pyro's source code to get the desired behaviour above. As a result every time we generate a new sample using SGHMC or SGLD we have to call `initialize_model` so that it gives us the correct $\tilde{U}(\theta)$ for some given minibatch $\tilde{\mathcal{D}}$, although this is a small price to pay for much quicker gradient computations.

4 Experiments

4.1 Simulated examples

In this part, we reproduce and extend the section "Simulated Scenarios" by writing our own python codes from scratch based on the Matlab version of the original paper [CFG14a]. A number of simulated experiments are conducted to validate theoretical results and demonstrate the differences between (1) standard HMC, (2) the naive implementation of stochastic gradient HMC, which simply replaces the gradient with a stochastic gradient, and (3) the proposed method incorporating friction, SGHMC. We also compare SGHMC to the first-order Langevin dynamics of SGLD.

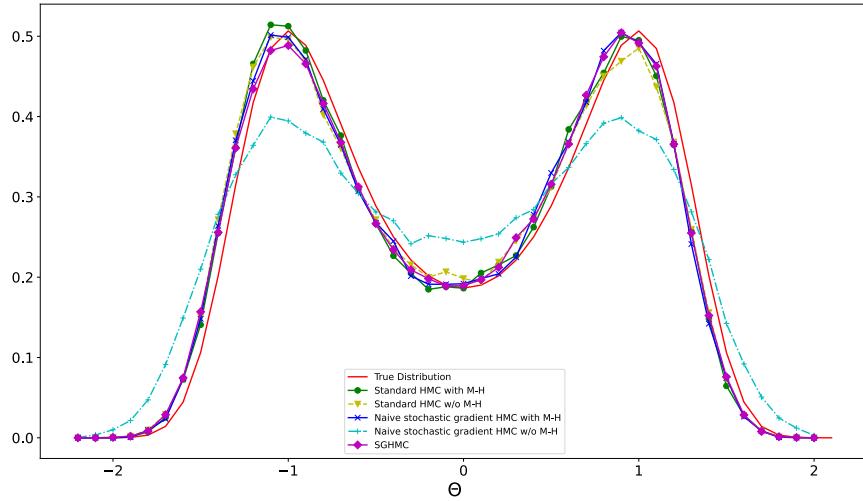
Firstly, to explore the performances of various sampling methods with exact gradients relative to stochastic gradients, we set up several simulated experiments based on the true target distribution with $U(\theta) = -2\theta^2 + \theta^4$ mentioned in original paper. Also, we test on $U(\theta) = \frac{1}{2}\theta^2$ as a simple extension which can help us confirm the conclusion here and also help us more thoroughly understand the Fig. 2(a) that has the same target distribution.

As a baseline, we consider the standard HMC methods, both with and without the MH correction. Then we compare with the naive stochastic HMC methods in Alg. 1 of Section 2.1, which simply replaces $\nabla U(\theta)$ with $\nabla \tilde{U}(\theta)$, both with and without as MH correction same as the above. Finally, we compare to the SGHMC, Alg. 2 of Section 2.3, without MH correction. Specifically, based on the current model parameters and sample size, 4 is chosen as the covariance of the stochastic gradient noise. Hence, $\nabla \tilde{U}(\theta) = \nabla U(\theta) + \mathcal{N}(0, 4)$ is used for stochastic gradient based samplers. And $\epsilon = 0.1$ is employed for all the cases. In all variants of HMC above, the momentum is resampled every 50 steps.

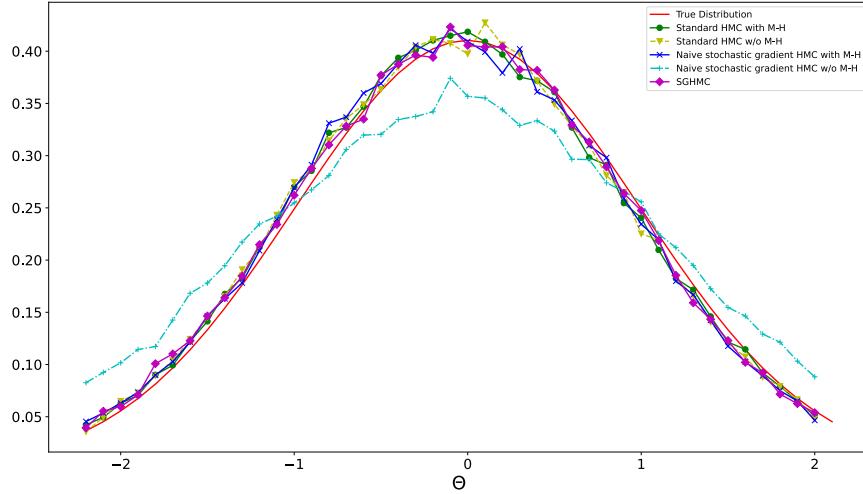
Compared with results in Fig. 1 of the original paper, we perfectly reproduce the simulations on the true distribution $U(\theta) = -2\theta^2 + \theta^4$ in Fig. 1(a). Also, we produce the tests on another true distribution $U(\theta) = \frac{1}{2}\theta^2$ as shown in Fig. 1(b). From both figures, we can see that both the HMC and SGHMC algorithms match perfectly with the target function even without an MH correction, which validates the theoretical results; that is, both standard HMC and SGHMC maintain π as the invariant distribution as $\epsilon \rightarrow 0$. However, the natural implementation of the stochastic approximation can be arbitrarily bad without MH step. This can be explained by the Corollary 3.1 in original paper. Because π is no longer invariant under the dynamics of naive stochastic gradient HMC, a correction step must be introduced even before considering errors introduced by the discretization of the dynamical system in practice.

As a conclusion, HMC with stochastic gradients requires a frequent costly MH correction step since the noise injected by the stochastic gradient no longer leads to Hamiltonian dynamics with the desired stationary target distribution. In contrary, SGHMC introduces a variant that uses second-order Langevin dynamics with a friction term that counteracts the effects of the noisy gradient, maintaining the desired target distribution as the invariant distribution.

Secondly, to simply simulating from the discretized Hamiltonian dynamical



(a) Distributions of different sampling algorithms for the target function $U(\theta) = -2\theta^2 + \theta^4$



(b) Distributions of different sampling algorithms for the target function $U(\theta) = \frac{1}{2}\theta^2$

Figure 1: Empirical distributions associated with various sampling algorithms relative to the different true target distributions.

systems with the various samplers compared, we compare their resulting trajectories of points (θ, r) . The compared methods are totally the same as the above. As for the details, discretization of various HMC methods are implemented in 15000 steps using the target distribution $U(\theta) = \frac{1}{2}\theta^2$ and ϵ is set as 0.1. For the noisy scenarios, the gradient is replaced by $\nabla \tilde{U}(\theta) = \nabla U(\theta) + \mathcal{N}(0, 4)$, the same as Fig. 1(b).

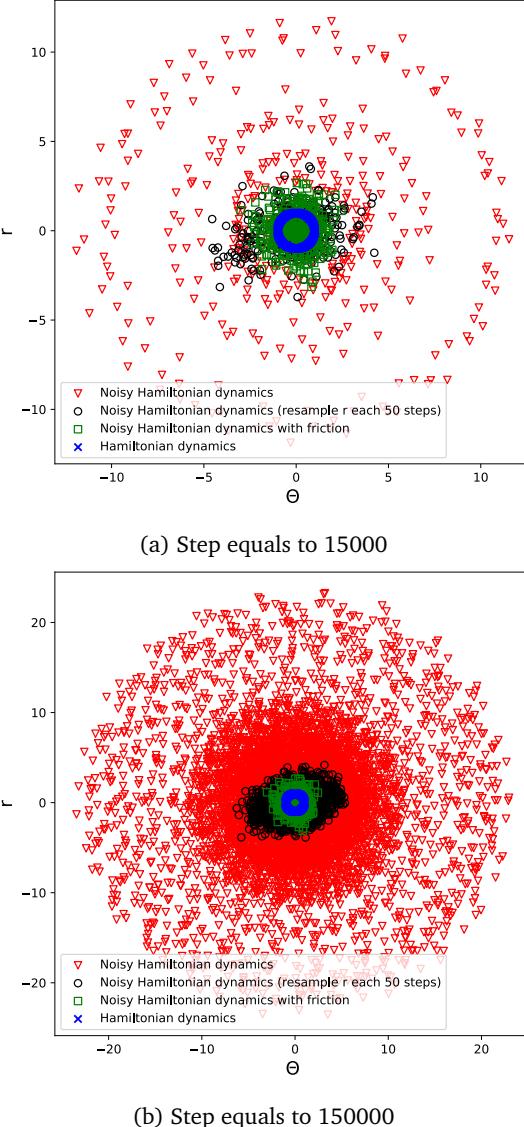


Figure 2: Points (θ, r) simulated from discretizations of various Hamiltonian dynamics of different steps using $U(\theta) = \frac{1}{2}\theta^2$.

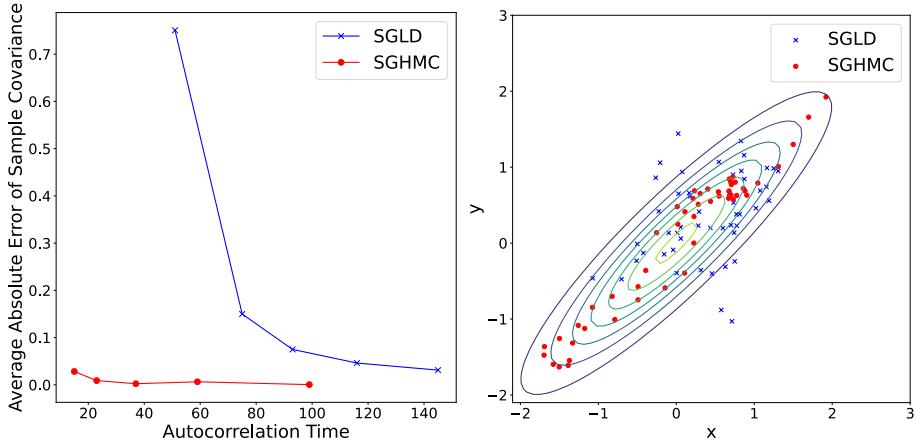
Compared with results in Fig. 2 of the original paper, we reproduce the Fig. 2(a) in a perfect way. And we find that the path of (θ, r) from the noisy Hamiltonian system without any friction or steps diverges significantly. To correct the divergence, the naive stochastic gradient HMC without MH steps can

be implemented through periodic resampling of the momentum, which may not yield the correct target distribution, as illustrated in Fig. 1(b). Then to correct this behaviour, SGHMC can be used to modify this dynamical system. The results can show that the straightforward modification to the Hamiltonian dynamics can alleviate the issues of the noise introduced by stochastic gradients. In particular, the modification allows us to again achieve the desired π as the invariant distribution of the continuous Hamiltonian dynamical system. Moreover, Fig. 2(a) can help explain Corollary 3.1 in the original paper by seeing that even finite-length simulations from the noisy system can diverge quite substantially from those of the noise-free system.

Additionally, we did some extensions on the steps setting for points simulation. As shown in Fig. 2(b), when steps are more than 15000 steps, like ten times or even more, the noisy Hamiltonian system will diverge more crazily, however, the systems with frictions or resamples stay more calm. This can well explained Theorem 3.1 in the original paper; that is the noise-free Hamiltonian dynamics preserve entropy, while the additional noise term strictly increases entropy. Then, jointly, the entropy strictly increases over time. This hints at the fact that the distribution $p_t(\theta, r)$ tends toward a uniform distribution, which can be very far from the target distribution π .

Thirdly, to prove the efficiency of HMC in sampling from correlated distributions, SGLD [WT11] is selected for comparing with SGHMC when sampling from a bivariate Gaussian with positive correlation. Specifically, $U(\theta) = \frac{1}{2}\theta^T \Sigma^{-1} \theta$, $\nabla U(\theta) = \Sigma^{-1} \theta + \mathcal{N}(0, I)$ with $\Sigma_{11} = \Sigma_{22} = 1$ and correlation $\rho = \Sigma_{12} = 0.9$. For each method, we run five different settings of the initial step size with a linearly decreasing scale and generate ten million samples. For one set per step-size setting, the autocorrelation time of the samples and the average absolute error of the resulting sample covariance are calculated .

Compared with results in Fig. 3(a) of the original paper, we excellently reproduce the Fig. 3(a). As the step size decreases, SGLD has a high autocorrelation time while SGHMC has very low autocorrelation times with even lower estimation error. In conclusion, SGHMC is indeed efficiently exploring the posterior distribution.



(a) Autocorrelation time versus estimation error for the five settings (b) First 50 samples of SGHMC and SGLD

Figure 3: Contrasting sampling of a bivariate Gaussian with correlation using SGHMC versus SGLD.

At last, to see more intuitively about the distributions of samples from SGHMC and SGLD, we reproduce Fig. 3(b) in original paper related to Fig. 3(b) here by showing the first 50 samples generated by the two samplers. We can conclude that SGLD’s random-walk behaviour makes it perform worse to explore the tails of the distribution. However, SGHMC instead can drive the sampler to move along the distribution contours all the way around.

4.2 Bayesian Neural Networks for Classification

For the Bayesian neural network (BNN) MNIST classification we actually used the reparameterisation of SGHMC described in the section “Connection to SGD with Momentum” of [CFG14b]. The SGHMC algorithm is reframed in terms of learning rate and momentum decay, and simulates the following dynamics instead:

$$\begin{cases} \nabla \theta = v \\ \nabla v = -\eta \nabla \tilde{U}(\theta) - \alpha v + \mathcal{N}(0, 2(\alpha - \hat{\beta})\eta) \end{cases}$$

where $\eta = \epsilon^2 M^{-1}$, $\alpha = \epsilon M^{-1} C$, $\hat{\beta} = \eta M^{-1} \hat{B}$. In all our experiments in this section we set the mass matrix M to the identity, and the noise model $\hat{\beta} = \hat{B} = 0$. Other than the architecture of the BNN there are now only 3 hyperparameters for SGHMC, the learning rate η , the momentum decay α and the batch size $|\tilde{\mathcal{D}}|$, in all our experiments we fixed $|\tilde{\mathcal{D}}| = 500$ which follows from [CFG14b].

To build on top of the work in [CFG14b] we implemented learning rate annealing for SGLD and following [WT11] we weighted the samples by the learning rate as follows:

$$\mathbb{E}[f(\theta)] \approx \frac{\sum_{t=1}^T \epsilon_t f(\theta_t)}{\sum_{t=1}^T \epsilon_t}$$

Although since f is a classifier we can ignore the denominator. Our BNN followed the same architecture as in [CFG14b], that is one linear with 100 hidden units followed by ReLU activation followed by another linear layer and a log softmax for multi-class classification. The weights and biases for both linear layers are sampled from univariate standard normal distributions, but the Pyro method `to_event()` declares dependence between the parameters.

Our implementations of SGD and SGD with momentum are meant to be used directly with Pyro, and so Gaussian priors on the weights and biases is equivalent to L2 regularization in the non-Bayesian paradigm. We experimented with regularization strengths of $\lambda \in \{0.1, 1.0, 10.0\}$ and found $\lambda = 1.0$ to be the most effective. Additionally we implemented weight decay for both SGD and SGD with momentum but found that this didn’t improve anything in this setting.

For the momentum based algorithms, SGHMC and SGD with momentum, we tried $\eta \in \{1.0, 2.0, 4.0, 8.0\} \times 10^{-6}$, and $\alpha \in \{0.1, 0.01, 0.001\}$. For SGHMC the best configuration was $\eta = 2.0 \times 10^{-6}$, $\alpha = 0.01$, and for SGD with momentum the best configuration was $\eta = 1.0 \times 10^{-6}$, $\alpha = 0.01$.

For SGLD and SGD, we tried $\eta \in \{1.0, 2.0, 4.0, 6.0\} \times 10^{-5}$, for SGLD we also tried learning rate annealing but it proved not to make much of a difference in this setting so we ignored it in the end. The best configuration for SGLD was $\eta = 4.0 \times 10^{-5}$, and for SGD the best configuration was $\eta = 1.0 \times 10^{-5}$.

For MNIST we ran each of the algorithms for 800 epochs with 50 warmup epochs. For the sampling algorithms we do Bayesian averaging over all the sampled parameterisations of the BNN after warmup as is described in Section II of [Jos+20] and report the test error. For the optimization algorithms we take most recent sample / set of parameters as a point estimate and report the test error. Figure 4 presents our results.

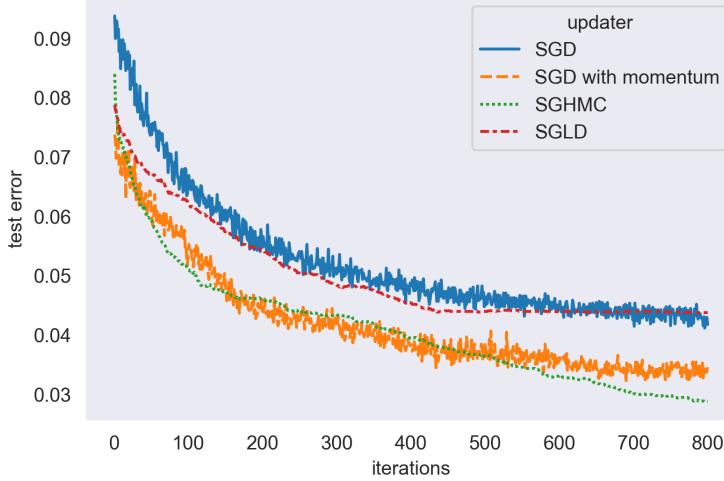


Figure 4: Reproducing the MNIST classification experiment from [CFG14b]; SGHMC ($\eta = 2.0 \times 10^{-6}, \alpha = 0.01, \text{resample_n} = 0$), SGLD ($\eta = 4.0 \times 10^{-5}$), SGD ($\eta = 1.0 \times 10^{-5}$), SGD with momentum ($\eta = 1.0 \times 10^{-6}, \alpha = 0.01$)

The results we get from MNIST classification align very closely with those in [CFG14b] and so we come to the same conclusion; the need for scalable and efficient Bayesian inference algorithms. The key benefit of BNNs is that we are not overconfident on out-of-distribution examples, Figure 5a illustrates that we still maintain this property when using SGHMC to approximately sample from the posterior distribution. We additionally conducted a brief comparison between Variational Inference (VI) and SGHMC in this setting, Figure 5b outlines our findings. The initial results suggest that SGHMC performs better than VI

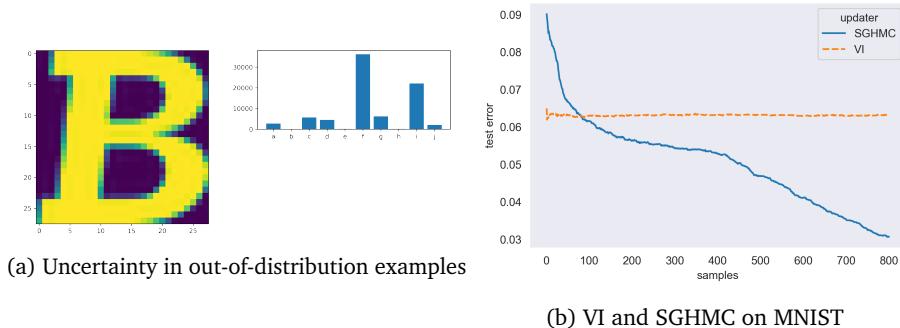


Figure 5: **Left** (a) illustrates that we get uncertainty estimates on out-of-distribution examples. **Right** (b) compares VI (Renyi ELBO, $\alpha = 0.01$, `num_particles = 2`) and SGHMC ($\eta = 2.0 \times 10^{-6}, \alpha = 0.01, \text{resample_n} = 0$) on MNIST. For VI we draw 80000 samples from the variational posterior q_ϕ and report the test error by Bayesian averaging. For SGHMC we do the same, except we are approximately sampling from the true posterior $p(\theta | D)$.

in this setting, although this is not the full picture. Once VI fits the variational posterior distribution q_ϕ as closely as possible to the true posterior it takes

only hundreds of samples to characterise q_ϕ , whereas SGHMC requires several more samples to characterise the true posterior. In practice storing thousands of parameterisations of the same NN is very costly and so this is probably why VI is a more popular choice for Bayesian inference.

We conclude this section by demonstrating our algorithms can be applied to other datasets and more complicated models, Figure 6a presents the results of running the same BNN architecture on FashionMNIST, and Figure 6b demonstrates that our implementation of SGHMC can be used with convolutional neural networks (CNNs).

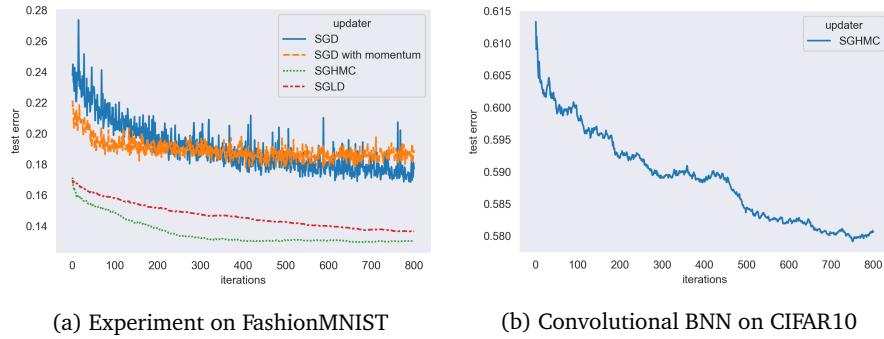


Figure 6: **Left** (a) FashionMNIST classification; SGHMC ($\eta = 1.0 \times 10^{-6}$, $\alpha = 0.01$, `resample_n = 0`), SGLD ($\eta = 1.0 \times 10^{-5}$), SGD ($\eta = 1.0 \times 10^{-5}$), SGD with momentum ($\eta = 1.0 \times 10^{-6}$, $\alpha = 0.01$); `warmup_epochs = 100`. **Right** (b) Convolutional BNN with 2 convolutional layers, batch norm, max pooling and tanh activations followed by two Bayesian linear layers with tanh activation; SGHMC ($\eta = 1.0 \times 10^{-6}$, $\alpha = 0.01$, `resample_n = 0`) ; `warmup_epochs = 150`.

5 NUTS

5.1 The Basic Algorithm

In our current description of the SGHMC algorithm we have the user-defined hyperparameter m , the number of steps iterated over before we take a sample. If this number is too small, our samples will be correlated, and hence successive samples would appear to follow a random walk, and we would get slow mixing times. We demonstrate this behaviour by training 3 versions of SGHMC with $m = 1, 3, 5$ (with ϵm fixed). The learning curves below in figure 7 show that increasing m increases the speed with which SGHMC reaches low error rates.

However, if m is too large we waste computational power, as we continue to step through time even though each sample is already independent of the last. We would like to set m to its optimal value.

Hoffman and Gelman introduce the algorithm NUTS in their paper “The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo” [HG11]. In its basic form this algorithm removes the need for a user to input a value for m in the standard HMC implementation. We converted this NUTS algorithm from being based on HMC to being based on SGHMC, and investigated its power. We will begin by presenting the high level overview of the original NUTS algorithm for HMC sampling.

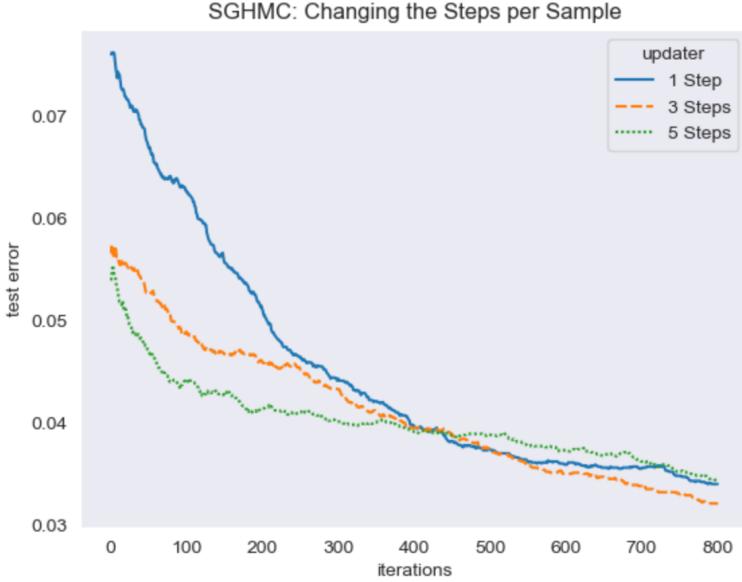


Figure 7: Changing the number of steps per sample. Each agent was run for 100 warmup epochs.

The key idea is the concept of a ‘U-Turn’. This is the point at which the samples of θ start to get closer to the initial value of θ , rather than away from it. This marks the point at which further steps will likely only waste computational power. Mathematically, for current position θ , initial position θ_0 and momentum r , this corresponds to the time at which:

$$\frac{d}{dt} |\theta - \theta_0|^2 = 2(\theta - \theta_0) \cdot \frac{d\theta}{dt} = 2(\theta - \theta_0) \cdot r < 0$$

where we use the fact that in the dynamics of HMC (and also SGHMC) we have $\frac{d\theta}{dt} = r$. This simple fact suggests an algorithm in which we draw the sample θ once we have performed enough steps so that $(\theta - \theta_0) \cdot r < 0$. However this is too simplistic an approach - as HMC is an instance of the Metropolis Hastings algorithm we require the Markov Chain of θ to be reversible, which is not the case here.

To remedy this problem, NUTS requires keeping track of a set \mathcal{B} , which contains all values of (θ, r) as steps are taken both forward and backwards in time. The values at the earliest and latest times considered across a single trajectory are labelled (θ^-, r^-) and (θ^+, r^+) respectively. NUTS starts from a single (θ, r) and then steps either forward or backwards one step. It then steps forward or backwards 2 steps, then 4 steps, then 8 steps etc. until a ‘U-Turn’ is seen. A description is as follows:

1. Start with an initial (θ_0, r_0)
2. Set $n \leftarrow 1$, $\mathcal{B} \leftarrow \{(\theta_0, r_0)\}$, $(\theta^-, r^-) \leftarrow (\theta_0, r_0)$ and $(\theta^+, r^+) \leftarrow (\theta_0, r_0)$
3. While there is no U-Turn at (θ^-, r^-) nor at (θ^+, r^+)
(ie $(\theta^+ - \theta^-) \cdot r^- \geq 0$ and $(\theta^+ - \theta^-) \cdot r^+ \geq 0$):
 - (a) Choose to go *forwards* or *backwards* in time, each with probability $\frac{1}{2}$.
 - (b) If *forwards*:

- i. $(\theta, r) \leftarrow (\theta^+, r^+)$.
 - ii. Repeat n times:
 - A. $(\theta, r) \leftarrow$ Step forward in time from (θ, r)
 - B. $\mathcal{B} \leftarrow (\theta, r) \cup \mathcal{B}$
 - iii. $(\theta^+, r^+) \leftarrow (\theta, r)$
 - (c) If *backwards*:
 - i. $(\theta, r) \leftarrow (\theta^-, r^-)$
 - ii. Repeat n times:
 - A. $(\theta, r) \leftarrow$ Step backwards in time from (θ, r)
 - B. $\mathcal{B} \leftarrow (\theta, r) \cup \mathcal{B}$
 - iii. $(\theta^-, r^-) \leftarrow (\theta, r)$
 - (d) $n \leftarrow 2n$
4. Carefully choose a subset $\mathcal{C} \subseteq \mathcal{B}$. This step is the key to the Markov Chain being reversible, however for clarity's sake we will not go into detail as to how this is done here.
5. Sample an element of \mathcal{C}

The benefit to this algorithm is that it removes the need to set the number of steps performed before we sample, as we keep stepping until a ‘U-Turn’ is seen. We should note here that in the original form of NUTS, the ‘step’ being referred to in *3biiA* and *3ciiA* is the step of the HMC algorithm. We edited the NUTS Pyro source code to make it perform SGHMC steps, and we named this SGNUTS.

There were some doubts as to whether the NUTS algorithm would work when using SGHMC steps instead of HMC steps. This was because NUTS requires the ability to step backwards in time, while an SGHMC step includes an injection of stochastic noise. As there is no action that undoes this injection, there was a worry that the backwards step through time in NUTS would become a problem. We explain how we attempted to solve this problem at the end of the next section.

5.2 Implementation of SGNUTS

To build SGNUTS we started with the Pyro source code for NUTS [Ube19] which takes the Pyro HMC class as its parent. We altered this to instead inherit from our SGHMC class. This required removing step-size adaptation and mass matrix adaptation functionality from NUTS, as our implementation of SGHMC was not able to interface with this. It also required introducing some caching methods into the SGHMC class - to help keep things simple in our original SGHMC class we did this in a new class, named SGHMC_for_NUTS. Most importantly, we had to alter the (θ, v) step update rule which is hardcoded in NUTS. This meant that instead of being the HMC update step it was now the SGHMC update step of:

$$\begin{aligned} \theta &\leftarrow \begin{cases} \theta + \epsilon M^{-1} r, & \text{forwards step} \\ \theta - \epsilon M^{-1} r, & \text{backwards step} \end{cases} \\ r &\leftarrow \begin{cases} r - \epsilon \nabla \tilde{U}(\theta) - \epsilon CM^{-1} r + \mathcal{N}(0, 2(C - \hat{B})\epsilon), & \text{forwards step} \\ r + \epsilon \nabla \tilde{U}(\theta) - \epsilon CM^{-1} r + \mathcal{N}(0, 2(C - \hat{B})\epsilon), & \text{backwards step} \end{cases} \end{aligned}$$

In particular, note that there is an injection of stochastic noise and momentum-reducing friction in both the forward steps and the backward steps. This is not ideal, as our backwards step in time does not undo a forward step, which likely breaks the reversibility of the Markov Chain being considered in NUTS. However we investigated this nonetheless.

5.3 Results

We tested the SGNUTS algorithm on the BNNs described earlier. We ran SGNUTS for only 20 warmup epoch and 50 epochs as the code was slower than SGHMC to run. The learning curves were as shown in Figure 8.

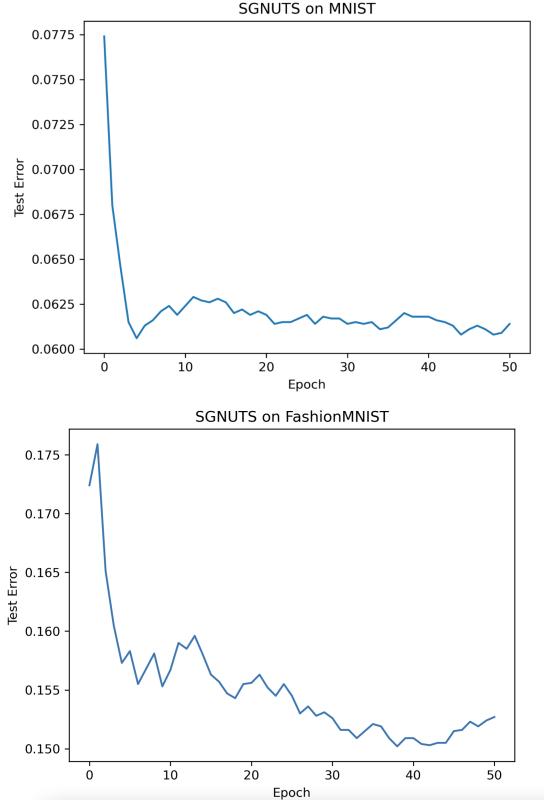


Figure 8: Classifying MNIST and FashionMNIST with SGNUTS

The final accuracies were 0.94 for MNIST and 0.85 for FashionMNIST which are similar to the accuracies obtained using SGHMC (0.97 and 0.86 respectively). This accuracy demonstrates that there is convergence to the true posterior in SGNUTS. This suggests that, like SGHMC itself, the SGNUTS Markov Chain is likely reversible in some (non traditional) sense. Denoting the posterior as $\pi(\theta, r)$, and transition kernels as $P(\theta, r | \theta', r')$, it is shown in [CFG14b] that SGHMC satisfies:

$$\pi(\theta, r)P_{SGHMC}(\theta, r | \theta', r') = \pi(\theta', -r')P_{SGHMC}(\theta', -r' | \theta, -r)$$

and it is suggested that this is the reason why the SGHMC algorithm works, despite not being reversible in the traditional sense. It would be interesting to consider if a similar property holds for P_{SGNUTS} .

While accurate, SGNUTS is slower to run than SGHMC to produce a single sample. It does however reach relatively high accuracies in a short amount of time. We measured the accuracy after the first epoch of SGHMC and NUTS as we changed the number of warmup epochs, and we measured how long they took to train.

Number of Warmup-Epochs	SGHMC		SGNUTS	
	Accuracy	Time (s)	Accuracy	Time (s)
0	0.7747	3.2	0.8715	28.3
1	0.5764	4.5	0.9296	127.2
2	0.4141	6.2	0.9231	453.3
5	0.6471	11.4		
10	0.7100	20.3		
25	0.8866	50.2		

In particular, SGNUTS reached an accuracy of 0.87 in 28s, while SGHMC reached 0.89 in 50s. This suggests a new algorithm that uses both SGNUTS and SGHMC — we could run a single epoch of SGNUTS so that we quickly approach the posterior distribution, at which point we switch to running SGHMC. It should be noted the speed up suggested by the above results would only be 30s seconds, however maybe on more complex datasets this could be larger. It would be interesting to investigate this if we had more time.

6 Conclusion

For this project, we have implemented a number of algorithms for the purpose of replicating and extending the experiments in [CFG14b]. We opted for a close integration with Pyro which allowed us to make use of its probabilistic programming framework.

Replicating the analysis in [CFG14b], we tested SGHMC on a BNN for classifying MNIST digits, comparing it to SGLD, SGD and SGD with momentum. We obtained similar results to the original experiment. Moreover, we expanded the analysis by testing SGMHC on a BNN for the Fashion-MNIST dataset and a CNN for CIFAR10. In both cases SGMHC performed very well. In the former we further compared SGLD, SGD and SGD with momentum, with SGMHC producing the superior performance. Given the close integration of our codebase with Pyro, it is easy to extend to further models and datasets, which we would do if given more time.

As a further extension, we briefly compared SGHMC with the more popular variational inference method, using our BNN on the MNIST dataset. Over a large number of epochs SGHMC performs better than VI. However, the latter converges faster, and requires fewer samples to give a representative picture of the posterior. With more time, we would like to investigate this further, for example by comparing the sampling efficiency more in-depth.

Our implementation of SGHMC provided the option of estimating the noise model using the observed information, as suggested in [CFG14b]. (More details based on results of experiments). Given more time, we would like to investigate alternative methods of estimating the noise model, ideally ones less computationally expensive.

Our last extension to [CFG14b] was the specification and implementation of the model method SGNUTS, which combines SGHMC with the No-U-Turn Sampler. In spite of the lack of reversibility of the corresponding Markov process, our algorithm performs well on a BNN for the MNIST and Fashion-MNIST datasets. We found that while SGNUTS takes longer than SGHMC to produce a single sample, but reaches high accuracies in a relatively small number of steps. If we were to continue this project, we would investigate in more detail the theoretical and practical aspects of SGNUTS.

References

- [AKW12] Sungjin Ahn, Anoop Korattikara and Max Welling. *Bayesian Posterior Sampling via Stochastic Gradient Fisher Scoring*. 2012. DOI: 10.48550/ARXIV.1206.6380. URL: <https://arxiv.org/abs/1206.6380>.
- [Bin+19] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall and Noah D. Goodman. ‘Pyro: Deep Universal Probabilistic Programming’. In: *J. Mach. Learn. Res.* 20.1 (Jan. 2019), pp. 973–978. ISSN: 1532-4435.
- [CFG14a] Tianqi Chen, Emily Fox and Carlos Guestrin. ‘Simulation Code’. In: 2014. URL: <https://github.com/tqchen/ML-SGHMC/tree/master>.
- [CFG14b] Tianqi Chen, Emily Fox and Carlos Guestrin. ‘Stochastic Gradient Hamiltonian Monte Carlo’. In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Beijing, China: PMLR, June 2014, pp. 1683–1691. URL: <https://proceedings.mlr.press/v32/cheni14.html>.
- [Den12] Li Deng. ‘The mnist database of handwritten digit images for machine learning research’. In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.
- [Dua+87] Simon Duane, A.D. Kennedy, Brian J. Pendleton and Duncan Roweth. ‘Hybrid Monte Carlo’. In: *Physics Letters B* 195.2 (1987), pp. 216–222. ISSN: 0370-2693. DOI: [https://doi.org/10.1016/0370-2693\(87\)91197-X](https://doi.org/10.1016/0370-2693(87)91197-X). URL: <https://www.sciencedirect.com/science/article/pii/037026938791197X>.
- [HG11] Matthew Hoffman and Andrew Gelman. ‘The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo’. In: 2011. URL: <https://arxiv.org/abs/1111.4246>.
- [Jos+20] Laurent Valentin Jospin, Wray Buntine, Farid Boussaid, Hamid Laga and Mohammed Bennamoun. ‘Hands-on Bayesian neural networks—a tutorial for deep learning users’. In: *arXiv preprint arXiv:2007.06823* (2020).
- [Kri09] Alex Krizhevsky. *Learning Multiple Layers of Features from Tiny Images*. 2009. URL: <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>.
- [Nea11] Radford M. Neal. ‘Handbook of Markov Chain Monte Carlo’. In: ed. by BySteve Brooks, Andrew Gelman, Galin Jones and Xiao-Li Meng. Chapman and Hall/CRC, 2011. Chap. MCMC Using Hamiltonian Dynamics, pp. 113–162. ISBN: 9780429138508.

- [Ube19] Inc. Uber Technologies. ‘Pyro NUTS Code’. In: 2019. URL: https://docs.pyro.ai/en/stable/_modules/pyro/infer/mcmc/nuts.html.
- [WT11] 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)*. Citeseer. 2011, pp. 681–688.
- [XRV17] Han Xiao, Kashif Rasul and Roland Vollgraf. *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms*. 2017. DOI: 10.48550/ARXIV.1708.07747. URL: <https://arxiv.org/abs/1708.07747>.