

# Incorporating the ChEES Criterion into Sequential Monte Carlo Samplers

Andrew Millard\*, Joshua Murphy\*, Daniel Frisch<sup>†</sup>, Simon Maskell\*

\*Dept. of EEE, University of Liverpool, United Kingdom

<sup>†</sup>Intelligent Sensor-Actuator-Systems Laboratory (ISAS), Karlsruhe Institute of Technology, Germany

Emails: amill212@liverpool.ac.uk, jmurph98@liverpool.ac.uk, daniel.frisch@kit.edu, smaskell@liverpool.ac.uk

**Abstract**—Markov chain Monte Carlo (MCMC) methods are a powerful but computationally expensive way of performing non-parametric Bayesian inference. MCMC proposals which utilise gradients, such as Hamiltonian Monte Carlo (HMC), can better explore the parameter space of interest if the additional hyper-parameters are chosen well. The No-U-Turn Sampler (NUTS) is a variant of HMC which is extremely effective at selecting these hyper-parameters but is slow to run and is not suited to GPU architectures. An alternative to NUTS, Change in the Estimator of the Expected Square HMC (ChEES-HMC) was shown not only to run faster than NUTS on GPU but also sample from posteriors more efficiently. Sequential Monte Carlo (SMC) samplers are another sampling method which instead output weighted samples from the posterior. They are very amenable to parallelisation and therefore being run on GPUs while having additional flexibility in their choice of proposal over MCMC. We incorporate (ChEES-HMC) as a proposal into SMC samplers and demonstrate competitive but faster performance than NUTS on a number of tasks.

## I. INTRODUCTION

Bayesian inference is a versatile way of making predictions and quantifying uncertainty while incorporating prior knowledge for a variety of applications such as deep learning [1], epidemiology [2], and environmental modelling [3].

Inference on more complex posteriors may require the use of sampling methods such as Markov chain Monte Carlo (MCMC) [4]. These sampling methods propose local moves within the distribution and build up a chain of particles which can be used to calculate statistics of functions on the distribution. MCMC can require many iterations to properly converge to a posterior so improved computational resources and tailoring the algorithm to exploit these architectures, often through parallelization, have contributed to its increased popularity [5]–[7]. Better proposals such as Hamiltonian Monte Carlo (HMC) make MCMC more efficient by using gradients to inform local moves but introduce tunable hyper-parameters which can be difficult to select [8]. The most popular HMC variant is the No-U-Turn Sampler (NUTS) which automatically tunes these hyper-parameters [9]. NUTS is the primary choice of sampler in numerous probabilistic programming languages like Stan, TensorFlow Probability and NumPyro [10]–[12]. However, the NUTS algorithm is not well-suited to take advantage of GPUs due to its complex control flow and inherent recursion [13], [14]. Change in the Estimator of the Expected Square HMC (ChEES-HMC) was proposed as an alternative adaptive HMC variant better suited

to GPU architectures and demonstrated significant speed-up while matching the performance of NUTS [15].

Sequential Monte Carlo (SMC) samplers are another method of sampling from complex posteriors by again using local moves to iteratively propagate a set of weighted samples around a distribution [16]. The importance sampling components of SMC samplers are easy to parallelise and mitigate some of the concerns of running a number of MCMC chains in parallel and combining them when they may not all have converged [17], [18].

Our contribution is to incorporate ChEES into the SMC framework and explore the effect of the choice of the quasi-random number generator used to jitter trajectory length.

The paper proceeds as follows: Section II gives an introduction to ChEES-HMC in the context of MCMC and Section III presents SMC and how ChEES is incorporated as a proposal. Section IV lays out the experiments with Section V providing results and discussion with conclusions and future work given in Section VI.

## II. MARKOV CHAIN MONTE CARLO

Consider the problem of sampling parameters  $\theta \in \mathbb{R}^D$  from a posterior distribution proportional to  $\pi(\theta) = p(\mathbf{x}|\theta)q_0(\theta)$  where  $p(\mathbf{x}|\theta)$  is a likelihood function and  $q_0(\theta)$  is a specified prior distribution over the parameters.

MCMC is a common general-purpose way of obtaining samples from intractable posteriors. The algorithm proceeds with a user-selected initial state,  $\theta_0$  and subsequently building up a chain of  $M$  parameter samples by proposing new samples according to a proposal distribution

$$\theta' \sim q(\cdot|\theta_{m-1}). \quad (1)$$

To ensure that samples come from the posterior distribution of interest, the sampler must be aperiodic, irreducible, obey detailed balance and the chain must leave the target distribution invariant. If the proposal is reversible, we can use the Metropolis-Hastings acceptance criterion and thus the samples come from the target [19] as our sampling process is now invariant.

$$\alpha(\theta_{m-1}, \theta') = \min \left( 1, \frac{\pi(\theta')q(\theta_{m-1}|\theta')}{\pi(\theta_{m-1})q(\theta'|\theta_{m-1})} \right). \quad (2)$$

A newly proposed sample,  $\theta'$  is accepted and added to the chain if a random variable,  $u \sim \mathcal{U}(0, 1)$ , drawn from a uniform distribution is less than the Metropolis-Hastings criterion in (2)

$$\theta_m = \begin{cases} \theta' & \text{if } u < \alpha(\theta_{m-1}, \theta'), \\ \theta_{m-1} & \text{otherwise.} \end{cases} \quad (3)$$

#### A. Hamiltonian Monte Carlo

HMC is a sampler which uses gradients to make more informed moves around the posterior [20]. It augments the posterior to be  $\pi(\theta, \mathbf{p})$  with a momentum variable usually taken from a Gaussian distribution,  $\mathbf{p} \sim \mathcal{N}(0, \mathcal{M})$ , where  $\mathcal{M}$  is the mass matrix and is typically set to an identity matrix  $\mathcal{M} = \mathbf{I}_D$ . The joint distribution can be written as

$$\pi(\theta, \mathbf{p}) = \exp\{-H(\theta, \mathbf{p})\}, \quad (4)$$

where,  $H(\cdot, \cdot)$ , the Hamiltonian is

$$H(\theta, \mathbf{p}) = -\log \pi(\mathbf{p}|\theta) - \log \pi(\theta) \quad (5)$$

$$= -\frac{1}{2}\|\mathbf{p}\|^2 - \log \pi(\theta). \quad (6)$$

which can be interpreted as a combination of kinetic and potential energy respectively. Therefore, the sample location and momentum can be updated using Hamilton's equations. These equations are generally intractable so a more practical approach is to use the leapfrog integrator to solve them numerically

$$\mathbf{p}_{l+\frac{1}{2}} = \mathbf{p}_l + \frac{\epsilon}{2}\nabla \log \pi(\theta_l), \quad (7)$$

$$\theta_{l+1} = \theta_l + \epsilon\mathbf{p}_{l+\frac{1}{2}}, \quad (8)$$

$$\mathbf{p}_{l+1} = \mathbf{p}_{l+\frac{1}{2}} + \frac{\epsilon}{2}\nabla \log \pi(\theta_{l+1}). \quad (9)$$

The leapfrog process is repeated for a certain number of user-specified leapfrog steps  $L$  with step size  $h$ . Pseudocode for the leapfrog algorithm can be found in Algorithm 1. Upon completion of  $L$  steps, leapfrog proposes a new sample and momentum, setting  $\theta' = \theta_L$  and  $\mathbf{p}' = \mathbf{p}_L$ , which are accepted according to the criterion

$$\alpha((\theta_{m-1}, \mathbf{p}_{m-1}), (\theta', \mathbf{p}')) = \min(1, \exp(-H(\theta', \mathbf{p}') + H(\theta_{m-1}, \mathbf{p}_{m-1}))). \quad (10)$$

---

#### Algorithm 1 Leapfrog Algorithm

---

**Require:** Initial state  $\theta$ , momentum  $\mathbf{p}$ , step size  $\epsilon$  and number of leapfrog steps  $L$

**for**  $l = 1$  to  $L$  **do**

    Half step update the momentum (7)

    Update the state (8)

    Complete the momentum update (9)

**end for**

**return**  $\theta_L, \mathbf{p}_L$

---

The step size  $\epsilon$  can be selected by numerous adaptive schemes such as dual-averaging [21], optimisation of the

expected squared jump distance [22], [23] and other adaptive MCMC methods. These do not generally interfere with the benefits of SMC and its parallelisation capacities [24]. However, selection of  $L$  is a little more challenging.

#### B. No-U-Turn Sampler

The most popular adaptive trajectory length selection algorithm is NUTS [9]. At each MCMC iteration, NUTS builds up a binary tree, using the leapfrog algorithm to build up trajectories in both directions alternately while doubling the number of leapfrog steps each time it considers switching direction. The construction of the tree stops when the trajectory makes a "U-turn",

$$(\theta^+ - \theta^-) \cdot \mathbf{p}^- < 0 \quad \text{or} \quad (\theta^+ - \theta^-) \cdot \mathbf{p}^+ < 0, \quad (11)$$

so that the tree ends up proposing just enough samples that some are far away from the point from which the tree is grown from.  $\theta^+$  and  $\theta^-$  represent the two furthest left and right points respectively in the binary tree. The trajectory is then randomly sampled to give  $\theta'$  which is accepted or rejected based on the criterion in (4).

Additional steps are taken to maintain detailed balance, such as requiring that the points within the binary tree meet four key conditions in order to be selected as a potential sample  $\theta'$ . The full details of these conditions can be found in the original paper [9].

#### C. Change in the Estimator of the Expected Square

Using the Change in the Estimator of the Expected Square (ChEES) criterion [15] is an effective alternative for the NUTS algorithm when adapting the trajectory length. The ChEES criterion is the following

$$\text{ChEES} = \frac{1}{4}\mathbb{E}[(\|\theta' - \mathbb{E}[\theta]\|^2 - \|\theta - \mathbb{E}[\theta]\|^2)^2] \quad (12)$$

This criterion promotes exploration by maximizing changes in sample variance, thereby reducing autocorrelation. For each chain  $c \in \{1, \dots, C\}$ , it is evaluated using proposed and previous states  $\theta^{(c)'}$ ,  $\theta_{m-1}^{(c)}$ , and their means  $\hat{\theta}'$ ,  $\hat{\theta}$ . The gradient estimate  $\hat{g}^{(c)}$  reflects how trajectory length influences sample dispersion, scaled by the jittered length  $l_m$  and momentum  $\mathbf{p}^{(c)}$ .

To enhance exploration, trajectory lengths are jittered via a Halton sequence  $h_m$ , with  $l_m = h_m L_{m-1}$  [25]. During warm-up, ChEES is maximized using gradient descent to adapt  $L_m$ , based on the weighted average gradient:

$$\hat{g} = \frac{\sum_c \alpha^{(c)} \hat{g}^{(c)}}{\sum_c \alpha^{(c)}} \quad (13)$$

This updates  $\log L_m$  via Adam, with a moving average  $\bar{L}$  maintained. After  $m = M_{\text{warmup}}$ ,  $L_m$  is fixed to  $\bar{L}$ .

Compared to NUTS, ChEES is often more efficient while still effectively adapting trajectory lengths. In SMC, multiple chains allow trajectory length to be tuned in parallel during warm-up. Algorithm 2 provides the full procedure.

---

**Algorithm 2** ChEES-HMC running on C chains.

---

**Require:** Initial state  $\theta_0^{(c)}$  for each chain  $c \in \{1, \dots, C\}$ , step size  $\epsilon$ , initial trajectory length  $L_0$ , desired number of samples  $M$ , number of adaptation steps  $M_{\text{warmup}}$ , random number sequence  $h_{1:M}$ .

- 1: Initialise moving averages  $\bar{L} = 0$ .
- 2: **for**  $m = 1$  to  $M$  **do**
- 3:   Sample momentum  $\mathbf{p}^{(c)} \sim \mathcal{N}(0, \mathcal{M})$ .
- 4:   Select jittered trajectory length  $l_m = h_m L_{m-1}$ .
- 5:   Propose new sample  $\theta^{(c)'}$  and momentum  $\mathbf{p}^{(c)'}$  using leapfrog( $\theta_{m-1}^{(c)}, \mathbf{p}^{(c)'}, \epsilon, \lceil l_m/\epsilon \rceil$ ).
- 6:   Compute acceptance probabilities  $\alpha^{(c)}$  using (10)
- 7:   Select  $\theta_m^{(c)}$  and  $\mathbf{p}_m^{(c)}$  according to (3)
- 8:   **if**  $m < M_{\text{warmup}}$  **then**
- 9:     Estimate the mean of the proposed and old states:  
10:      $\hat{\theta}' = \frac{1}{C} \sum_c \theta^{(c)'}$ ,  $\hat{\theta} = \frac{1}{C} \sum_c \theta_{m-1}^{(c)}$
- 11:     Compute trajectory gradient estimates:  
12:      $\hat{g}^{(c)} = l_m \left( \|\theta^{(c)'} - \hat{\theta}'\|^2 - \|\theta_{m-1}^{(c)} - \hat{\theta}\|^2 \right) (\theta^{(c)'} - \hat{\theta}')^\top \mathbf{p}^{(c)}$
- 13:     Update log-trajectory length  $\log L_m$  with Adam using weighted gradient:  
14:      $\hat{g} = \frac{\sum_c \alpha^{(c)} \hat{g}^{(c)}}{\sum_c \alpha^{(c)}}$
- 15:     Update moving averages trajectory  $\bar{L} \leftarrow 0.9\bar{L} + 0.1L_m$
- 16:     **end if**
- 17:     **if**  $m = M_{\text{warmup}}$  **then**
- 18:        $L_{m:M} \leftarrow \bar{L}$
- 19:     **end if**
- 20: **end for**

---

### III. SEQUENTIAL MONTE CARLO

SMC is another algorithm for targeting static posterior distributions via  $K$  sequential importance sampling steps and resampling when necessary [16]. The joint distribution of all states until  $k = K$  is defined as

$$\pi(\theta_{1:K}) = \pi(\theta_K) \prod_{k=1}^K L(\theta_{k-1}|\theta_k), \quad (14)$$

where  $L(\theta_{k-1}|\theta_k)$  is the L-kernel, which is a user-defined probability distribution. The choice of this distribution can greatly impact the efficacy of the sampler [26].

At  $k = 1$ ,  $J$  samples  $j = 1, \dots, J$  are drawn from a prior distribution  $q_0(\cdot)$  as follows:

$$\theta_0^j \sim q_0(\cdot), \quad \forall j, \quad (15)$$

and weighted according to

$$\mathbf{w}_1^j = \frac{\pi(\theta_0^j)}{q_0(\theta_0^j)}, \quad \forall j. \quad (16)$$

At  $k > 1$ , subsequent samples are proposed based on samples from the previous iteration via a proposal distribution,  $q(\theta_k^j|\theta_{k-1}^j)$  by

$$\theta_k^j \sim q(\cdot|\theta_{k-1}^j). \quad (17)$$

---

**Algorithm 3** SMC sampler running for  $K$  iterations and  $J$  samples.

---

Sample  $\{\theta_0^{(j)}\}_{j=1}^J \sim q_0(\cdot)$   
Set initial weights  $\mathbf{w}_0^j$  using (16)  
**for**  $k = 1$  to  $K$  **do**  
  **for**  $j = 1$  to  $J$  **do**  
    Normalise weights using (20)  
  **end for**  
  Calculate  $J_{\text{eff}}$  using (19)  
  **if**  $J_{\text{eff}} < J/2$  **then**  
    Resample  $[\theta_k^1 \dots \theta_k^J]$  with probability  $[\tilde{\mathbf{w}}_k^1 \dots \tilde{\mathbf{w}}_k^J]$   
    Reset all weights to  $\frac{1}{J}$   
  **end if**  
  **for**  $j = 1$  to  $J$  **do**  
    Propagate samples  $\theta_{k-1}^j$  according to (17)  
    Update sample weights  $\mathbf{w}_k^j$  using (18)  
  **end for**  
**end for**

---

These samples are then weighted according to

$$\mathbf{w}_k^j = \mathbf{w}_{k-1}^j \frac{\pi(\theta_k^j)}{\pi(\theta_{k-1}^j)} \frac{L(\theta_{k-1}^j|\theta_k^j)}{q(\theta_{k-1}^j|\theta_k^j)}, \quad \forall i. \quad (18)$$

SMC samplers compute the Effective Sample Size (ESS) as a measure of the efficiency of the sampler at iteration  $k$  by

$$J^{\text{eff}} = \frac{1}{\sum_{j=1}^J \left( \tilde{\mathbf{w}}_k^j \right)^2}, \quad (19)$$

using the sum of the normalised weights which are calculated from

$$\tilde{\mathbf{w}}_k^j = \frac{\mathbf{w}_k^j}{\sum_{j=1}^J \mathbf{w}_k^j}, \quad \forall j. \quad (20)$$

As iterations continue, one weight tends to dominate which is known as particle degeneracy and can be mitigated by resampling. Resampling is undertaken if  $J^{\text{eff}} < J/2$ . There are a variety of potential resampling schemes [27] including the optimally parallelised systematic resampling schemes outlined in [28], [29]. Here we utilise multinomial resampling for ease of implementation. Samples are assigned an unnormalised weight of  $\frac{1}{J}$  after resampling.

The weighted samples can be used to picture the whole distribution as well as realise estimates of the expectations of functions on the distribution through

$$\begin{aligned} \mathbb{E}_\pi[f(\theta_k)] &= \int f(\theta_k) \pi(\theta_k) d\theta_k \\ &= \int f(\theta_k) \frac{\pi(\theta_k)}{q(\theta_k)} q(\theta_k) d\theta_k \approx \tilde{\mathbf{f}}_k \end{aligned} \quad (21)$$

$$\tilde{\mathbf{f}}_k = \sum_{j=1}^J \tilde{\mathbf{w}}_k^j f(\theta_k^j), \quad (22)$$

Pseudocode for a generic SMC sampler can be found in Algorithm 3.

### A. Proposals in SMC

A common but naive choice of the proposal distribution in (17) is a Gaussian with a mean of  $\theta_{k-1}^j$  and a covariance of  $\Sigma \in \mathbb{R}^{D \times D}$ , such that

$$q(\theta_k^j | \theta_{k-1}^j) = \mathcal{N}(\theta_k^j; \theta_{k-1}^j, \Sigma), \quad \forall j. \quad (23)$$

This is also referred to as a random walk proposal. Gradient-based proposals originating from MCMC like Langevin [30], HMC [31], [32] and NUTS [33] have been effectively incorporated into SMC. MCMC proposals can be included in SMC with or without an accept-reject step [24], [33]. Here we choose not to have an accept-reject step when including ChEES as a proposal in SMC (SMC-ChEES) and comparing to NUTS in SMC.

A sub-optimal but easily implementable approach to selecting the L-kernel in (18) is to choose the same distribution as the forwards proposal

$$L(\theta_{k-1}^j | \theta_k^j) = q(\theta_{k-1}^j | \theta_k^j), \quad \forall i, \quad (24)$$

With gradient-based proposals both proposal and L-kernel can be evaluated in terms of the stochastic momentum component  $\mathbf{p}$  [33], [34]

$$q(\theta_k^j | \theta_{k-1}^j) = \mathcal{N}(\mathbf{p}_{k-1}; 0, \mathcal{M}) \left| \frac{df_{\text{LF}}(\theta_{k-1}, \mathbf{p}_{k-1})}{d\mathbf{p}_{k-1}} \right|^{-1}, \quad (25)$$

$$L(\theta_{k-1}^j | \theta_k^j) = \mathcal{N}(-\mathbf{p}_k; 0, \mathcal{M}) \left| \frac{df_{\text{LF}}(\theta_k, -\mathbf{p}_k)}{d\mathbf{p}_k} \right|^{-1}. \quad (26)$$

where  $f_{\text{LF}}$  is the leapfrog process. When using (25) and (26) their Jacobians cancel in (18). We also apply this change of variable to the proposal and L-kernel when we evaluate SMC-ChEES.

### B. SMC-ChEES

SMC-ChEES is an extension of the HMC proposals discussed in section 2, as it provides a principled alternative trajectory adaption technique to NUTS. When using this in an SMC context, we can use  $J$  random seeds to jitter each trajectory length by a different amount at each iteration, and then use this to compute the acceptance rate  $\alpha$  needed for the optimization procedure to maximize ChEES during the warm up phase. Even after warm up, although we fix  $L$  we can still use different jitters for each sample to effectively explore the distribution.

1) *Random Number Generators (RNGs)*: In [15] the random number sequence  $h_{1:M}$  was generated from a 1-dimensional Halton sequence [25]. We explore the usage of the following random and quasi-random number generation schemes to generate a matrix of  $J \times K$  random numbers  $h_{1:JK}$  for use with ChEES as a proposal in an SMC sampler.

- 1) No jitter: all  $J \times K$  numbers are set to 1.
- 2) Uniform random: the random number sequence is drawn from a standard uniform

$$h_{jk} \sim U(0, 1). \quad (27)$$

- 3) N-d Halton:

$$h_{jk} = \sum_{d=0}^{\infty} \text{digits}_d^k(j) k^{-d-1} \quad (28)$$

where  $\text{digits}_d(j)$  is the  $d^{\text{th}}$  digit of  $j$  represented in base- $k$  with the order of the digits reversed.

- 4) N-d Inverse Halton: as in (28) but the matrix is then sorted in reverse order of  $k$  to ensure lower discrepancy bases are used at the end of the sampling sequence.
- 5) 1-d Halton:  $J \times K$  numbers are taken from (28) with base set to 2.
- 6) N-d Primes:

$$h_{jk} = j\sqrt{\mathbb{P}_k} \pmod 1 \quad (29)$$

where  $\mathbb{P}_k$  is the  $k^{\text{th}}$  prime number.

- 7) N-d Inverse Primes: as in (29) with the matrix sorted in reverse order of  $k$ .
- 8) 1-d Golden Ratio:

$$\frac{((j-1)K + k)(\sqrt{5} - 1)}{2} \pmod 1. \quad (30)$$

- 9) Equidistant:  $J$  points are created from  $h_j = j/J$ . These  $h_j$  points are shuffled  $K$  times to fill the  $h_{jk}$  matrix.
- 10) Offset Equidistant: as with Equidistant but each number is also perturbed with a draw from a  $U(0, 0.1)$ .
- 11) N-d Sobol:

$$h_{jk} = \text{digits}_1^2(j)\nu_1^k \oplus \text{digits}_2^2(j)\nu_2^k \oplus \dots \quad (31)$$

where  $\nu_d^k$  are direction numbers typically obtained as the coefficients of a primitive polynomial. We use the scipy implementation which takes direction numbers from [35].

- 12) Inverse N-d Sobol: as in (31) with the matrix sorted in reverse order of  $k$ .
- 13) 1-d Sobol:  $J \times K$  numbers are taken from (28) with  $k = 1$ .

## IV. EXPERIMENTAL SET-UP

In this section we present the distributions we sample from for evaluation of our method. In all examples the SMC samplers have  $J = 1000$  particles and are run for  $K = 200$  iterations, the first 100 of which are taken as burn-in. Initial samples are drawn from a prior of appropriate dimensionality for the target  $\theta_0 \sim \mathcal{N}(\mathbf{0}_D, \mathbf{I}_D)$ . ChEES is initialised with  $L = 5$ . Each sampler is run 10 times and an average of estimates and metrics is reported.

### A. Gaussian

The first example is a multivariate Gaussian with  $D = 5$  and parameters

$$\theta \sim \mathcal{N}(\mu, \Sigma) \quad (32)$$

$$\mu = [-4, -2, 0, 2, 4]^T \quad (33)$$

$$\Sigma = \text{diag}(1, 1.5, 2, 2.5, 3) \quad (34)$$

A step size of  $\epsilon = 0.1$  was utilised.

	Gaussian		Ill-conditioned Gaussian		Banana		German Credit	
	$\nabla\text{eval}/N$	$J^{\text{eff}}/\nabla\text{eval}$	$\nabla\text{eval}/N$	$J^{\text{eff}}/\nabla\text{eval}$	$\nabla\text{eval}/N$	$J^{\text{eff}}/\nabla\text{eval}$	$\nabla\text{eval}/N$	$J^{\text{eff}}/\nabla\text{eval}$
<b>NUTS</b>	63.95	1.56e-02	1771.93	4.16e-04	468.58	2.12e-03	1952.19	3.61e-04
<b>No Jitter</b>	12.65	7.91e-02	501.00	1.63e-03	25.84	3.87e-02	501.00	1.40e-03
<b>1-d Uniform</b>	6.55	1.53e-01	501.00	1.27e-03	57.10	1.74e-02	143.48	5.11e-03
<b>N-d Halton</b>	11.02	9.08e-02	501.00	1.25e-03	73.80	1.35e-02	52.66	1.46e-02
<b>N-d Inverse Halton</b>	17.86	5.60e-02	501.00	1.21e-03	34.73	2.87e-02	72.77	1.04e-02
<b>1-d Halton</b>	8.01	1.25e-01	2.01	3.20e-01	21.96	4.55e-02	3.58	2.33e-01
<b>N-d Primes</b>	3.59	2.78e-01	501.00	1.20e-03	43.50	2.29e-02	102.15	7.50e-03
<b>N-d Inverse Primes</b>	5.55	1.80e-01	501.00	1.35e-03	50.34	1.98e-02	96.48	8.72e-03
<b>1-d Golden Ratio</b>	5.23	1.91e-01	501.00	1.26e-03	61.37	1.62e-02	92.92	9.00e-03
<b>N-d Equidistant</b>	6.46	1.55e-01	501.00	1.25e-03	52.61	1.89e-02	106.21	7.55e-03
<b>N-d Offset Equidistant</b>	9.06	1.10e-01	501.00	1.22e-03	52.63	1.89e-02	186.85	4.08e-03
<b>N-d Sobol</b>	7.28	1.37e-01	501.00	1.24e-03	62.26	1.60e-02	147.55	5.45e-03
<b>N-d Inverse Sobol</b>	8.68	1.15e-01	501.00	1.24e-03	77.89	1.28e-02	101.76	7.32e-03
<b>1-D Sobol</b>	8.76	1.14e-01	2.47	2.64e-01	34.43	2.90e-02	4.62	1.96e-01

TABLE I: Number of gradient evaluations per sample (smaller is better) and effective sample size per gradient evaluation (larger is better) averaged across iterations.

### B. Ill-conditioned Gauss

Another multivariate Gaussian with  $D = 100$ . A random orthogonal matrix  $\mathbf{Q} \in \mathbb{R}^{100 \times 100}$  is drawn uniformly from the Haar measure on the orthogonal group  $O(100)$ , ensuring that  $\mathbf{Q}\mathbf{Q}^\top = \mathbf{I}_{100}$ . The eigenvalues  $\{\lambda_i\}_{i=1}^{100}$  of the covariance matrix  $\Sigma$  are drawn from a Gamma distribution with shape parameter 0.5 and scale parameter 1. The condition number of  $\Sigma \approx 1.3 \times 10^5$  [15]. A step size of  $\epsilon = 0.001$  was used.

$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \quad (35)$$

$$\boldsymbol{\mu} = [0, \dots, 0]^\top \in \mathbb{R}^{100} \quad (36)$$

$$\lambda_j \sim \text{Gamma}(0.5, 1), \quad j = 1, 2, \dots, 100 \quad (37)$$

$$\mathbf{Q} \sim \text{Haar}(O(100)) \quad (38)$$

$$\Sigma = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top. \quad (39)$$

### C. Rosenbrock Distribution

The Rosenbrock (banana) distribution with  $D = 2$ . The joint distribution is a product of the marginals in dimension  $\boldsymbol{\theta}_{(1)}$  and  $\boldsymbol{\theta}_{(2)}$

$$\boldsymbol{\theta}_{(1)} \sim \mathcal{N}(0, 10) \quad (40)$$

$$\boldsymbol{\theta}_{(2)} \sim \mathcal{N}(0.03(\boldsymbol{\theta}_{(1)}^2 - 100), 1) \quad (41)$$

A step size of  $\epsilon = 0.01$  was used.

### D. German Credit

A logistic regression on the numerical German credit dataset [36]. Here  $D = 25$ . A step size of  $\epsilon = 0.001$  was used.

$$\mathbf{y}_n \sim \text{Bernoulli}(\sigma(\boldsymbol{\theta}^\top \mathbf{x}_n)) \quad (42)$$

$$\sigma(x) \triangleq \frac{1}{1 + e^{-x}} \quad (43)$$

## V. RESULTS

Fig. 1 shows the number of effective samples per gradient evaluation ( $J^{\text{eff}}/\nabla\text{eval}$ ) for each of the 4 experiments. It's clear that ChEES has a vastly greater  $J^{\text{eff}}/\nabla\text{eval}$  than NUTS across all tasks with every RNG. This is reflected in Table I

which shows the average number of gradient evaluations per sample and the average  $J^{\text{eff}}/\nabla\text{eval}$  across iterations. For the Gaussian, Ill-conditioned Gaussian and Banana targets, NUTS had approximately 4 to 5 times more average  $\nabla\text{eval}$  per particle than the ChEES methods and the same difference can be found in the average  $J^{\text{eff}}/\nabla\text{eval}$ .

A far larger disparity in the number of  $\nabla\text{eval}$  per particle and likewise in the  $J^{\text{eff}}/\nabla\text{eval}$  can be seen in the German Credit task where NUTS used about 80 times more evaluations than some of the ChEES methods. In our implementation, the maximum tree depth of NUTS was set to  $2^{11}$  meaning that no further leapfrog steps were taken beyond that number, even if the No-U-turn criterion in (11) was not met. NUTS took an average of 1952.19 ( $\nabla\text{eval}/N - 1$ ) leapfrog steps as there is also a gradient calculation undertaken prior to the leapfrog steps. This is far greater than on any other task and the maximum tree depth limit was frequently reached during the sampling process. The ChEES proposals, are less effected by their use on real world data and don't see the same spike in  $\nabla\text{eval}/N$ .

However, we do notice that on the ill-conditioned Gaussian example for nearly every method, we reach the maximum number of leapfrog steps manually set in the implementation. Therefore, we would need to increase the maximum trajectory length to in order to evaluate the effective sample size better. Despite reaching this limit though, all methods still produced good MSE on both the mean and variance showing that a longer trajectory length is potentially unnecessary. This also outlines that  $J^{\text{eff}}/\nabla\text{eval}$  evaluation, although a good metric for understanding computational efficiency, should not be used as a singular metric as we notice this does not always translate to producing meaningful samples from the posterior. For example in the 5-D example 2, we see that the No Jitter method has a relatively good  $J^{\text{eff}}/\nabla\text{eval}$ , but a suboptimal MSE.

The RNG that gives the best  $J^{\text{eff}}/\nabla\text{eval}$  and MSE results overall seems to be the 1-d Halton method which was employed in the original ChEES [15] paper while the 1-d Sobol method also produces good results on the same metrics. However this again demonstrates the need to consider

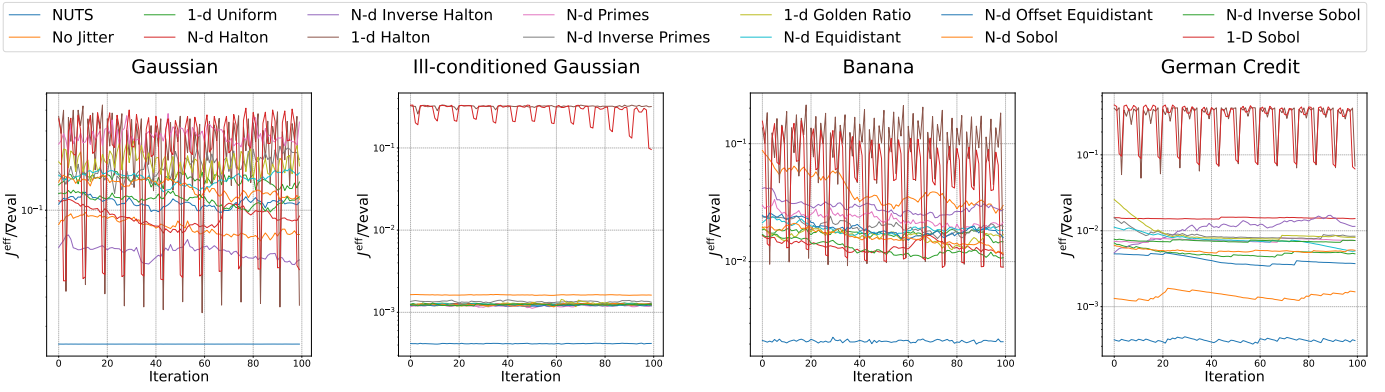


Fig. 1: Number of effective samples per gradient evaluation per iteration for the four experiments.

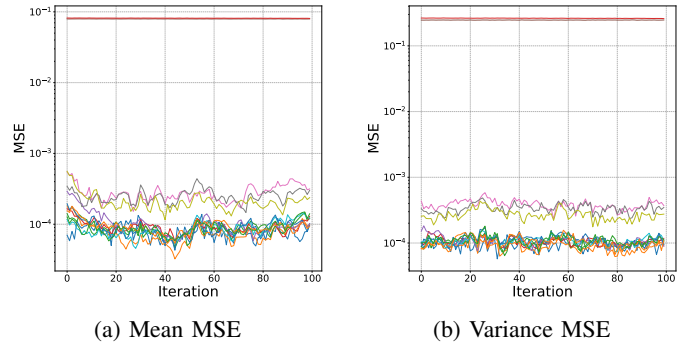
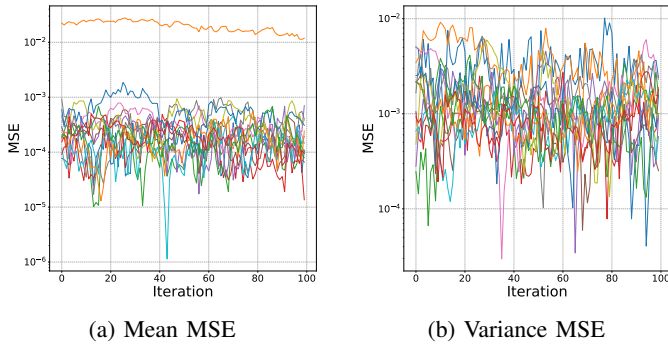


Fig. 2: MSE of the mean and variance estimates for the 5-dimensional Gaussian distribution obtained by the different proposals.

Fig. 3: MSE of the mean and variance estimates for the 100-dimensional Ill-conditioned Gaussian distribution obtained by the different proposals.

additional metrics as the performance of both RNGs on the Ill-conditioned Gauss and German Credit task are notably poorer.

In Fig. 1 the methods alternate for which has the highest  $J^{\text{eff}}/\nabla\text{eval}$  across iterations and this is also reflected by Table I. Our results support previous findings that jittering the trajectory length is helpful in drawing good samples from the posterior distribution [8].

Fig. 2 and 3 show the mean square error (MSE) for the mean and variance of the Gaussian and Ill-conditioned Gaussian. The mean and variance are realised using (22) and an average is then taken over the dimensions. NUTS is clearly the first to converge and the ChEES proposals converge to a similar MSE to NUTS within 15 iterations. In the Banana distribution NUTS explores furthest into the tails as shown in Fig. 4 which shows the positions of all particles across the last twenty iterations. Once again NUTS is the best proposal on the German Credit logistic regression task but the difference to the ChEES methods is marginal and therefore we may see another method produce marginally better results with more/different starting seeds.

## VI. CONCLUSIONS

In this paper we incorporate ChEES as a proposal into an SMC framework with a change of variables L-kernel and

explore the use of different RNG options to jitter the trajectory length. We demonstrate that ChEES is far more efficient in terms of the number of effective samples it achieves per gradient evaluation with this difference being particularly pronounced on the real-world German Credit logistic regression and Banana distribution examples.

We have also investigated the different RNG methods for jittering outperform no-jittering within the ChEES algorithm with the best overall performance coming from the 1-d Halton and 1-d Sobol methods. We note that the depth has been reached for many methods on the Ill-conditioned Gaussian example which requires more investigation. Despite hitting the maximum number of leapfrog steps, the MSE performs is still on par with NUTS.

Further work could focus on including step size adaption methods (such as dual averaging) into SMC-ChEES to minimise the number of hyper-parameters that need to be manually tuned. As SMC-ChEES picks a singular trajectory length which it then jitters, it is competitive with NUTS in scenarios where NUTS trajectory lengths may not differ significantly. Therefore, it would be useful to investigate ChEES on example problems where this is not the case, such as the funnel distribution [37].

SMC-ChEES could also be further evaluated on more real-

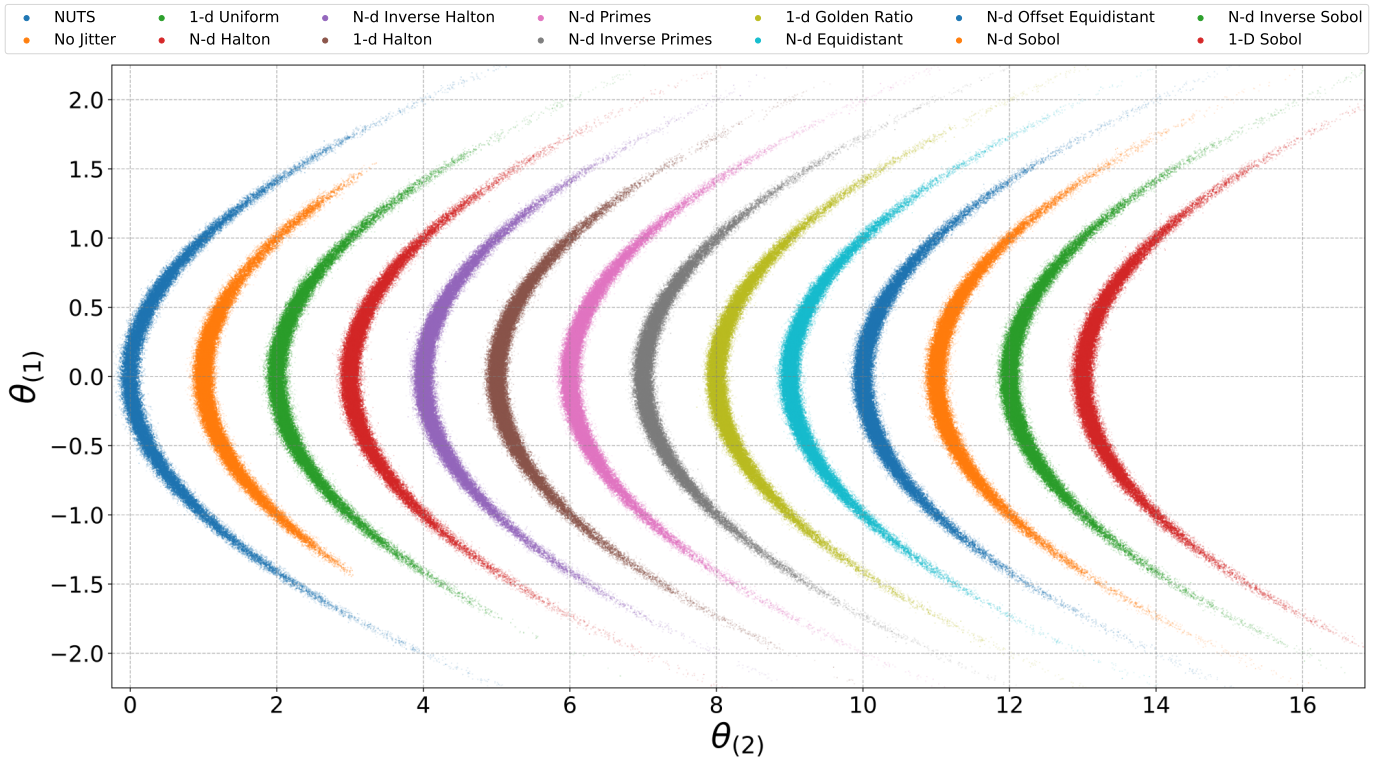


Fig. 4: Position of all samples for the last 20 iterations of the banana distribution.

	Accuracy	Precision	Recall	F1 Score	Specificity	AUROC
<b>NUTS</b>	0.78	0.66	0.53	0.58	0.89	0.71
<b>No Jitter</b>	0.78	0.66	0.53	0.58	0.89	0.71
<b>1-d Uniform</b>	0.76	0.64	0.42	0.51	0.90	0.66
<b>N-d Halton</b>	0.76	0.63	0.46	0.53	0.89	0.67
<b>N-d Inverse Halton</b>	0.77	0.65	0.44	0.53	0.90	0.67
<b>1-d Halton</b>	0.76	0.65	0.41	0.50	0.91	0.66
<b>N-d Primes</b>	0.77	0.65	0.44	0.53	0.90	0.67
<b>N-d Inverse Primes</b>	0.78	0.68	0.47	0.56	0.91	0.69
<b>1-d Golden Ratio</b>	0.77	0.64	0.47	0.54	0.89	0.68
<b>N-d Equidistant</b>	0.76	0.64	0.42	0.51	0.90	0.66
<b>N-d Offset Equidistant</b>	0.77	0.66	0.46	0.54	0.90	0.68
<b>N-d Sobol</b>	0.77	0.65	0.44	0.53	0.90	0.67
<b>N-d Inverse Sobol</b>	0.77	0.65	0.44	0.53	0.90	0.67
<b>1-D Sobol</b>	0.76	0.67	0.37	0.48	0.92	0.65

TABLE II: Logistic Regression Results for the German Credit Dataset (larger is better).

world data to see if the benefit of using alternative RNG methods with ChEES persists across other tasks.

#### ACKNOWLEDGMENT

AM and JM were funded by a Research Studentship jointly funded by the EPSRC Centre for Doctoral Training in Distributed Algorithms EP/S023445/1. This work was funded by Dstl in collaboration with the Royal Academy of Engineering via "Dstl-RAEng Research Chair in Information Fusion" under task RQ0000040616. SM thanks Dstl, UK MOD and the Royal Academy of Engineering for supporting this work. The views and conclusions contained in this paper are of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the UK MOD or the UK Government.

#### REFERENCES

- [1] R. M. Neal, *Bayesian Learning for Neural Networks*, vol. 118. Springer Science & Business Media, 2012.
- [2] D. A. Rasmussen, O. Ratmann, and K. Koelle, "Inference for Nonlinear Epidemiological Models Using Genealogies and Time Series," *PLoS Computational Biology*, vol. 7, no. 8, p. e1002136, 2011.
- [3] G. Papaefthymiou and B. Klockl, "MCMC for Wind Power Simulation," *IEEE transactions on energy conversion*, vol. 23, no. 1, pp. 234–240, 2008.
- [4] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, "An Introduction to MCMC for Machine Learning," *Machine learning*, vol. 50, pp. 5–43, 2003.
- [5] C. P. Robert, V. Elvira, N. Tawn, and C. Wu, "Accelerating MCMC Algorithms," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 10, no. 5, p. e1435, 2018.
- [6] A. S. Mahani and M. T. Sharabiani, "SIMD Parallel MCMC Sampling with Applications for Big-Data Bayesian Analytics," *Computational Statistics & Data Analysis*, vol. 88, pp. 75–99, 2015.

- [7] G. Mingas, L. Bottolo, and C.-S. Bouganis, "Particle MCMC Algorithms and Architectures for Accelerating Inference in State-Space Models," *International Journal of Approximate Reasoning*, vol. 83, pp. 413–433, 2017.
- [8] R. M. Neal, "MCMC using Hamiltonian Dynamics," *arXiv preprint arXiv:1206.1901*, 2012.
- [9] M. D. Hoffman, A. Gelman, *et al.*, "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1593–1623, 2014.
- [10] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A Probabilistic Programming Language," *Journal of statistical software*, vol. 76, 2017.
- [11] J. V. Dillon, I. Langmore, D. Tran, E. Brevdo, S. Vasudevan, D. Moore, B. Patton, A. Alemi, M. Hoffman, and R. A. Saurous, "TensorFlow Distributions," 2017.
- [12] D. Phan, N. Pradhan, and M. Jankowiak, "Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro," 2019.
- [13] A. Radul, B. Patton, D. Maclaurin, M. D. Hoffman, and R. A. Saurous, "Automatically Batching Control-Intensive Programs for Modern Accelerators," 2020.
- [14] J. Lao, C. Suter, I. Langmore, C. Chimisov, A. Saxena, P. Sountsov, D. Moore, R. A. Saurous, M. D. Hoffman, and J. V. Dillon, "tfp.mcmc: Modern markov chain monte carlo tools built for modern hardware," 2020.
- [15] M. Hoffman, A. Radul, and P. Sountsov, "An Adaptive-MCMC Scheme for Setting Trajectory Lengths in Hamiltonian Monte Carlo," in *International Conference on Artificial Intelligence and Statistics*, pp. 3907–3915, PMLR, 2021.
- [16] P. Del Moral, A. Doucet, and A. Jasra, "Sequential Monte Carlo Samplers," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 68, no. 3, pp. 411–436, 2006.
- [17] E. Drouiotis, A. Varsi, A. M. Phillips, S. Maskell, and P. G. Spirakis, "A Massively Parallel SMC Sampler for Decision Trees," *Algorithms*, vol. 18, no. 1, p. 14, 2025.
- [18] A. Varsi, L. Kekempanos, J. Thiyagalingam, and S. Maskell, "A Single SMC Sampler on MPI that Outperforms a Single MCMC Sampler," 2019.
- [19] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.
- [20] M. Betancourt, "A Conceptual Introduction to Hamiltonian Monte Carlo," *arXiv preprint arXiv:1701.02434*, 2017.
- [21] Y. Nesterov, "Primal-Dual Subgradient Methods for Convex Problems," *Mathematical programming*, vol. 120, no. 1, pp. 221–259, 2009.
- [22] C. Pasarica and A. Gelman, "Adaptively Scaling the Metropolis Algorithm using Expected Squared Jumped Distance," *Statistica Sinica*, pp. 343–364, 2010.
- [23] Z. Wang, S. Mohamed, and N. Freitas, "Adaptive Hamiltonian and Riemann Manifold Monte Carlo," in *International conference on machine learning*, pp. 1462–1470, PMLR, 2013.
- [24] A. Buchholz, N. Chopin, and P. E. Jacob, "Adaptive Tuning of Hamiltonian Monte Carlo within Sequential Monte Carlo," *Bayesian Analysis*, vol. 16, no. 3, pp. 745–771, 2021.
- [25] J. H. Halton, "On the Efficiency of Certain Quasi-Random Requences of Points in Evaluating Multi-Dimensional Integrals," *Numerische Mathematik*, vol. 2, pp. 84–90, 1960.
- [26] P. L. Green, L. Devlin, R. E. Moore, R. J. Jackson, J. Li, and S. Maskell, "Increasing the Efficiency of Sequential Monte Carlo Samplers through the use of Approximately Optimal L-Kernels," *Mechanical Systems and Signal Processing*, vol. 162, p. 108028, 2022.
- [27] R. Douc and O. Cappé, "Comparison of Resampling Schemes for Particle Filtering," in *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.*, pp. 64–69, Ieee, 2005.
- [28] A. Varsi, S. Maskell, and P. G. Spirakis, "An  $O(\log^2 N)$  Fully-Balanced Resampling Algorithm for Particle Filters on Distributed Memory Architectures," *Algorithms*, vol. 14, no. 12, p. 342, 2021.
- [29] C. Rosato, A. Varsi, J. Murphy, and S. Maskell, "An  $O(\log^2 N)$  SMC 2 Algorithm on Distributed Memory with an Approx. Optimal L-Kernel," in *2023 IEEE Symposium Sensor Data Fusion and International Conference on Multisensor Fusion and Integration (SDF-MFI)*, pp. 1–8, IEEE, 2023.
- [30] A. Sim, S. Filippi, and M. P. Stumpf, "Information Geometry and Sequential Monte Carlo," *arXiv preprint arXiv:1212.0764*, 2012.
- [31] D. Gunawan, R. Kohn, and M. N. Tran, "Robust Particle Density Tempering for State Space Models," *arXiv preprint arXiv:1805.00649*, 2018.
- [32] R. Daviet, "Inference with Hamiltonian Sequential Monte Carlo Simulators," *arXiv preprint arXiv:1812.07978*, 2018.
- [33] Devlin, Lee and Carter, Matthew and Horridge, Paul and Green, Peter L. and Maskell, Simon, "The No-U-Turn sampler as a proposal distribution in a sequential monte carlo sampler without accept/reject," *IEEE Signal Processing Letters*, vol. 31, pp. 1089–1093, 2024.
- [34] C. Rosato, J. Murphy, A. Varsi, P. Horridge, and S. Maskell, "Enhanced SMC2: Leveraging Gradient Information from Differentiable Particle Filters Within Langevin Proposals," *arXiv preprint arXiv:2407.17296*, 2024.
- [35] S. Joe and F. Y. Kuo, "Constructing Sobol Sequences with Better Two-Dimensional Projections," *SIAM Journal on Scientific Computing*, vol. 30, no. 5, pp. 2635–2654, 2008.
- [36] H. Hofmann, "Statlog (German Credit Data)." UCI Machine Learning Repository, 1994. DOI: <https://doi.org/10.24432/C5NC77>.
- [37] R. M. Neal, "Slice Sampling," *The annals of statistics*, vol. 31, no. 3, pp. 705–767, 2003.