IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 11, 2022, accepted March 17, 2022, date of publication March 22, 2022, date of current version April 1, 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3161443*

# Shadow Magnetic Hamiltonian Monte Carlo

**WILSON TSAKANE MONGWE**[1,2], **RENDANI MBUVHA**[3], **AND TSHILIDZI MARWALA**[1,2], **(Senior Member, IEEE)**
[1]School of Electrical Engineering, University of Johannesburg, Auckland Park 2000, South Africa
[2]Johannesburg Machine Intelligence Laboratory, Auckland Park 2006, South Africa
[3]School of Statistics and Actuarial Science, University of Witwatersrand, Johannesburg, Johannesburg 2006, South Africa

Corresponding author: Wilson Tsakane Mongwe (wilsonmongwe@gmail.com)

The work of Wilson Tsakane Mongwe and Rendani Mbuvha was supported by the Google Ph.D. Fellowships in Machine Learning. The work of Tshilidzi Marwala was supported by the National Research Foundation of South Africa.

**ABSTRACT** Incorporating partial momentum refreshment into Magnetic Hamiltonian Monte Carlo (MHMC) to create Magnetic Hamiltonian Monte Carlo with partial momentum refreshment (PMHMC) has been shown to improve the sampling performance of MHMC significantly. At the same time, sampling from an integrator-dependent shadow or modified target density has been utilised to boost the acceptance rates of Hamiltonian Monte Carlo (HMC), which leads to more efficient sampling as the integrator is better conserved by the shadow Hamiltonian than the true Hamiltonian. Sampling from the shadow Hamiltonian associated with the numerical integrator used in MHMC is yet to be explored in the literature. This work aims to address this gap in the literature by combining the benefits of the non-canonical Hamiltonian dynamics of MHMC with those achieved by targeting the modified Hamiltonian. We first determine the modified Hamiltonian associated with the MHMC integrator and use this to construct a novel method, which we refer to as Shadow Magnetic Hamiltonian Monte Carlo (SMHMC), that leads to better sampling behaviour when compared to MHMC while leaving the target distribution invariant. The new SMHMC method is compared to MHMC and PMHMC across various target posterior distributions, including datasets modeled using Bayesian Neural Networks and Bayesian Logistic Regression models.

**INDEX TERMS** Bayesian neural networks, Bayesian logistic regression, Hamiltonian Monte Carlo, magnetic Hamiltonian Monte Carlo, shadow Hamiltonian, Markov chain Monte Carlo.

## I. INTRODUCTION

Markov Chain Monte Carlo (MCMC) methods are a vital inference tool for probabilistic machine learning models [1]–[6]. MCMC algorithms are preferable to variational approaches [7], [8] as they are assured to converge to the correct target distribution if the sample size is adequately large [9], [10]. These methods are premised on constructing a Markov chain of samples that asymptotically, as the number of generated samples tends to infinity, converge to the desired equilibrium distribution. By definition, the samples generated by MCMC algorithms are auto-correlated, which means that they will have higher variance than classical Monte Carlo techniques. This branch of inference techniques was initially developed by physicists, with famous examples being the Metropolis-Hastings [11] method of Metropolis and Hastings and the now popular and go-to Hamiltonian Monte Carlo (HMC) algorithm of Duane *et al.* [12]. These MCMC methods then later entered the field of computational statistics, where they are used today to sample from various complex probabilistic models [13].

HMC has been enhanced over the last decade with some examples of the improved algorithms being Riemannian Hamiltonian Monte Carlo [2] which considers the local geometry of the target posterior to better explore the density, the No-U-Turn Sampler (NUTS) [3] which automatically tunes the trajectory length and step size parameters of HMC, Quantum-Inspired Hamiltonian Monte Carlo [14] which uses a random mass matrix as inspired by the behaviour of quantum particles, continuously-tempered Hamiltonian Monte Carlo [15], [16] which is suited for sampling from multi-modal distributions and also produces the Bayesian evidence metric which can be utilised for model comparison, Magnetic Hamiltonian Monte Carlo [4], [17] which employs non-canonical Hamiltonian dynamics to better explore the target posterior, as well as methods that use shadow Hamiltonians, such as Separable Shadow Hamiltonian Hybrid Monte Carlo [18] and Shadow Manifold

The associate editor coordinating the review of this manuscript and approving it for publication was Sajid Ali.

Hamiltonian Monte Carlo [5], to sample from high dimensional target densities while maintaining high sample acceptance rates [5], [18]–[23].

Magnetic Hamiltonian Monte Carlo (MHMC) utilises non-canonical dynamics to skillfully probe the target posterior distribution [4], [6], [20]. MHMC introduces a magnetic field to HMC which leads to reduced auto-correlations and efficient convergence [4], [24]. MHMC has been extended to manifolds by Brofos and Lederman [25] and shows good improvement over MHMC. Mongwe *et al.* [26] present a method for the automatic tuning of the step size and trajectory length parameters in MHMC, which shows improvement over MHMC. This method is based on the NUTS methodology of Hoffman and Gelman [3]. MHMC has the disadvantage that the magnetic component has to be specified by the user. In the existing literature on MHMC, there are no automated means of tuning the magnetic component [4], [6], [20].

It has been previously confirmed that the performance of HMC suffers from the deterioration in acceptance rates due to numerical integration errors as the system size increases [5], [19]. As MHMC is an extension of HMC, and becomes HMC when the magnetic component is absent, one would expect MHMC also suffers from the pathology of deterioration of acceptance rates as the system size increases. This deterioration in acceptance rates results in large auto-correlations between the generated samples, thus requiring large sample sizes. The decline of the acceptance rates can be reduced by using more accurate higher-order integrators, by using smaller step sizes, or by employing shadow Hamiltonians [23]. These first two approaches tend to be more computationally expensive than the latter approach [5], [23]. In this work, we explore the method of utilising shadow Hamiltonians.

Shadow Hamiltonian-based samplers have been successfully employed to manage the deterioration of sample acceptance as the system size and step sizes increases and lead to a more efficient sampling of the target posterior [18], [19], [22]. The shadow Hamiltonians are constructed by performing backward error analysis of the integrator and, as a result, are better preserved when compared to the true Hamiltonian [27]. Numerous strategies have been proposed for sampling from shadow Hamiltonians of diverse numerical integrators [5], [18], [19], [22], [28].

Mongwe *et al.* [29] introduce the Quantum-Inspired Magnetic Hamiltonian Monte Carlo algorithm. Their work explored the utility of employing a random mass matrix for the auxiliary momentum variable in MHMC, which is consistent with the behaviour of quantum particles. The results showed a significant improvement in sampling results when compared to MHMC. Magnetic Hamiltonian Monte Carlo with partial momentum refreshment (PMHMC) was introduced by Mongwe *et al.* [17] and shows that retaining some of the chains' past dynamics can improve the sampling performance of MHMC. The disadvantage of the above two approaches is the need to manually tune the volatility-of-volatility and momentum refreshment parameters that the authors introduce. Although these works addressed important gaps in the literature, they did not consider the potential benefits of sampling from the shadow Hamiltonian instead of the true Hamiltonian in MHMC. Mongwe *et al.* [28] explored the benefits of combining shadow Hamiltonians and partial momentum refreshment for the Separable Shadow Hamiltonian Monte Carlo method with good results. Sampling from the shadow Hamiltonian associated with the leapfrog-like integrator used in MHMC is yet to be explored in the literature.

In this work, we address this gap in the literature by deriving the fourth-order shadow Hamiltonian corresponding to the leapfrog-like integrator in MHMC. From this, we construct the novel Shadow Magnetic Hamiltonian Monte Carlo (SMHMC) algorithm, which leaves the target density invariant. We compare the performance of the proposed method to MHMC and the PMHMC algorithm in [17]. The empirical results on a multivariate Gaussian distribution with a dimension of ten, real-world benchmark datasets modeled using Bayesian Logistic Regression and Bayesian Neural Network targets show that SMHMC achieves higher effective sample sizes when compared to the other MCMC algorithms considered. The proposed method does, however, consume more computational resources than the other MCMC methods due to the requirement to compute the shadow Hamiltonian and thus leads to poor performance on a time-normalised basis. This is a crucial area of improvement of the proposed method, which we intend to address in the future.

### A. CONTRIBUTIONS

Our contributions in this work can be summarised as follows:

- A shadow Hamiltonian that is preserved up to fourth-order by the numerical integrator used in MHMC is derived.
- We combine the benefits of non-canonical Hamiltonian dynamics with the proprieties of shadow Hamiltonians to form the Shadow Magnetic Hamiltonian Monte Carlo (SMHMC) algorithm, which is guaranteed to leave the target density invariant.
- Numerical experiments across diverse targets show that the new algorithm outperforms the other considered MCMC techniques on an effective sample size basis.

The remainder of this paper is structured as follows: Sections II and III discuss the material that forms the basis of the new method, Section IV introduces the new method, Section V outlines the experiments conducted and Section VI discusses the results of the experiments. We then provide the conclusion in Section VII.

## II. MAGNETIC HAMILTONIAN MONTE CARLO

The Hamiltonian Monte Carlo (HMC) algorithm is composed of two steps: 1) the molecular dynamics step and 2) the Monte Carlo step. The molecular dynamics step involves integrating Hamiltonian dynamics, while the Monte Carlo step employs

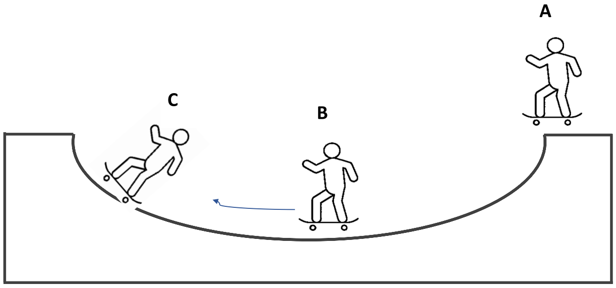

**FIGURE 1.** Illustration of the conservation of the Hamiltonian (i.e., total energy) through time as the skater moves from position A to C.

the Metropolis-Hastings (MH) algorithm to account for any errors introduced by the numerical integrator used in the molecular dynamics step [11], [12], [19], [22], [30]–[32]. Note that if we could exactly solve the molecular dynamics step, we would not need the Monte Carlo step.

HMC improves upon the MH [11] algorithm by utilising first-order gradient information of the unnormalised target posterior. This gradient information is used to guide HMC's exploration of the parameter space [12], [32]. This necessities that the target posterior function is differentiable and has support almost everywhere on $\mathbb{R}^D$, which is the case for the majority of machine learning models of interest. In HMC, the position vector $\mathbf{w}$ is augmented with auxiliary momentum variable $\mathbf{p}$, which is typically chosen to be independent of $\mathbf{w}$. The Hamiltonian $H(\mathbf{w}, \mathbf{p})$, which represents the total energy, from this system is written as follows:

$$H(\mathbf{w}, \mathbf{p}) = U(\mathbf{w}) + K(\mathbf{p}) \tag{1}$$

where $U(\mathbf{w})$ is potential energy or the negative log-likelihood of the target posterior distribution and $K(\mathbf{p})$ is the kinetic energy defined by the kernel of a Gaussian with a covariance matrix $\mathbf{M}$ [1]:

$$K(\mathbf{p}) = \frac{1}{2}\log\left((2\pi)^D|\mathbf{M}|\right) + \frac{\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p}}{2}. \tag{2}$$

Within this framework, the evolution of the physical system is governed by Hamiltonian dynamics [1], [33]. As the particle moves through time using Hamiltonian dynamics, the total energy is conserved over the entire trajectory of the particle, with kinetic energy being exchanged for potential energy and vice versa to ensure that the Hamiltonian or total energy is conserved [1], [33]. As an illustration, consider a person skating from position A to C as displayed in Figure 1. At position A, the person only has potential energy and no kinetic energy, and they only have kinetic energy at point B. At position C, they have both kinetic and potential energy. Throughout the movement from A to C, the total energy will be conserved if the individual traverses the space using Hamiltonian dynamics. This allows the individual to traverse long distances. This energy conservation property of Hamiltonian dynamics is key to the efficiency of HMC in exploring the target posterior.

The equations governing the Hamiltonian dynamics are defined by Hamilton's equations in a fictitious time $t$ as follows [31]:

$$\frac{d\mathbf{w}}{\partial t} = \frac{\partial H(\mathbf{w}, \mathbf{p})}{\partial \mathbf{p}}; \quad \frac{d\mathbf{p}}{\partial t} = -\frac{\partial H(\mathbf{w}, \mathbf{p})}{\partial \mathbf{w}}. \tag{3}$$

which can also be re-expressed as:

$$\frac{d}{\partial t}\begin{bmatrix}\mathrm{w}\\ \mathbf{p}\end{bmatrix} = \begin{bmatrix}\mathbf{0} & \mathbf{I}\\ -\mathbf{I} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\nabla_w \mathrm{H}(\mathrm{w}, \mathrm{p})\\ \nabla_p \mathrm{H}(\mathrm{w}, \mathrm{p})\end{bmatrix} \tag{4}$$

The Hamiltonian dynamics satisfy the following important properties, which make it ideal for efficiently generating distant proposals [30], [34]:

1) **Conservation of energy:** That is, the change of the Hamiltonian through time is zero as illustrated in Figure 1. Mathematically:

$$\begin{aligned}\frac{\partial H(\mathbf{w}, \mathbf{p})}{\partial t} &= \frac{\partial H(\mathbf{w}, \mathbf{p})}{\partial \mathbf{w}}\frac{\partial \mathbf{w}}{\partial t} + \frac{\partial H(\mathbf{w}, \mathbf{p})}{\partial \mathbf{p}}\frac{\partial \mathbf{p}}{\partial t}\\ &= \frac{\partial H(\mathbf{w}, \mathbf{p})}{\partial \mathbf{w}}\left(\frac{\partial H(\mathbf{w}, \mathbf{p})}{\partial \mathbf{p}}\right)\\ &\quad + \frac{\partial H(\mathbf{w}, \mathbf{p})}{\partial \mathbf{p}}\left(-\frac{\partial H(\mathbf{w}, \mathbf{p})}{\partial \mathbf{w}}\right)\\ &\Longrightarrow \frac{\partial H(\mathbf{w}, \mathbf{p})}{\partial t} = 0\end{aligned} \tag{5}$$

2) **Reversibility:** That is, the dynamics can be moved forward in time by a certain amount and backwards in time by the same amount to get back to the original position. Mathematically: Let $\Phi_{t,H}\begin{bmatrix}\mathbf{w_0}\\ \mathbf{p_0}\end{bmatrix}$ be the unique solution at time $t$ of equation (3) with initial position $\begin{bmatrix}\mathbf{w_0}\\ \mathbf{p_0}\end{bmatrix}$. As the Hamiltonian in equation (1) is time-homogeneous, we have that :

$$\begin{aligned}\Phi_{t,H}\circ\Phi_{s,H}\begin{bmatrix}\mathbf{w_0}\\ \mathbf{p_0}\end{bmatrix} &= \Phi_{t+s,H}\begin{bmatrix}\mathbf{w_0}\\ \mathbf{p_0}\end{bmatrix}\\ &\Longrightarrow \Phi_{-t,H}\circ\Phi_{t,H}\begin{bmatrix}\mathbf{w_0}\\ \mathbf{p_0}\end{bmatrix} = \begin{bmatrix}\mathbf{w_0}\\ \mathbf{p_0}\end{bmatrix}\end{aligned} \tag{6}$$

3) **Volume preservation:** This property serves to simplify the MH step in HMC so that it does not require a Jacobian term, as volume preservation means that the Jacobian term is equal to one [1], [35]. There have also been extensions of HMC that do not preserve volume [33].

These three properties are significant in that conservation of energy allows one to determine if the approximated trajectory is diverging from the expected dynamics, reversibility of the Hamiltonian dynamics ensures reversibility of the sampler, and volume preservation simplifies the MH acceptance step [1], [5].

The differential equation in equations (3) and (4) cannot be solved analytically in most instances. This necessitates the use of a numerical integration scheme. As the Hamiltonian in equation (1) is separable, to traverse the space, we can employ the leapfrog integrator [12], [31]. The position and momentum update equations for the leapfrog integration scheme are:

$$\begin{aligned} \mathbf{p}_{t+\frac{\epsilon}{2}} &= \mathbf{p}_t + \frac{\epsilon}{2}\frac{\partial H\left(\mathbf{w}_t, \mathbf{p}_t\right)}{\partial \mathbf{w}} \\ \mathbf{w}_{t+\epsilon} &= \mathbf{w}_t + \epsilon \mathbf{M}^{-1}\mathbf{p}_{t+\frac{\epsilon}{2}} \\ \mathbf{p}_{t+\epsilon} &= \mathbf{p}_{t+\frac{\epsilon}{2}} + \frac{\epsilon}{2}\frac{\partial H\left(\mathbf{w}_{t+\epsilon}, \mathbf{p}_{t+\frac{\epsilon}{2}}\right)}{\partial \mathbf{w}}. \end{aligned} \tag{7}$$

**Algorithm 1:** Hamiltonian Monte Carlo Algorithm

```
    Input: $N, \epsilon, L, w_{\text{init}}, H(w, p)$
    Output: $(w)_{m=0}^{N}$
1:  $w_0 \leftarrow w_{\text{init}}$
2:  for $m \to 1$ to $N$ do
3:      $p_{m-1} \sim \mathcal{N}(0, \mathbf{M})$
4:      $p_m, w_m = $ Leapfrog($p_{m-1}, w_{m-1}, \epsilon, L, H$) in equation (7)
5:      $\delta H = H(w_{m-1}, p_{m-1}) - H(w_m, p_m)$
6:      $\alpha_m = \min(1, \exp(\delta H))$
7:      $u_m \sim \text{Unif}(0, 1)$
8:      $w_m = $ Metropolis($\alpha_m, u_m, w_m, w_{m-1}$) in equation (8)
9:  end for
```

Due to the discretisation errors arising from the numerical integration, the Monte Carlo step in HMC utilises the MH algorithm in which the parameters $\mathbf{w}^*$ proposed by the molecular dynamics step are accepted with probability:

$$\text{P(accept } \mathbf{w}^*) = \min\left(1, \frac{\exp\left(-H(\mathbf{w}^*, \mathbf{p}^*)\right)}{\exp\left(-H(\mathbf{w}, \mathbf{p})\right)}\right). \tag{8}$$

Algorithm 1 shows the pseudo-code for the HMC where $\epsilon$ is the discretisation step size and $L$ is the trajectory length. The overall HMC sampling process follows a Gibbs sampling scheme, where we *fully* sample the momentum (see line 3 in Algorithm 1) and then sample a new set of parameters given the drawn momentum.

Magnetic Hamiltonian Monte Carlo (MHMC) is a special case of non-canonical HMC using a symplectic structure corresponding to motion of a particle in a magnetic field [4], [6]. MHMC extends HMC by endowing it with a magnetic field, which results in non-canonical Hamiltonian dynamics [4]. This magnetic field offers a significant amount of flexibility over HMC and encourages more efficient exploration of the posterior, which results in faster convergence and lower auto-correlations in the generated samples [4], [20], [24]. MHMC uses the same Hamiltonian as in HMC, but exploits non-canonical Hamiltonian dynamics where the canonical matrix now has a non-zero element on the diagonal. The MHMC dynamics are given as:

$$\frac{d}{\partial t}\begin{bmatrix}\mathbf{w} \\ \mathbf{p}\end{bmatrix} = \begin{bmatrix}\mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{G}\end{bmatrix}\begin{bmatrix}\nabla_w \mathrm{H(w, p)} \\ \nabla_p \mathrm{H(w, p)}\end{bmatrix} \tag{9}$$

where $\mathbf{G}$ is a skew-symmetric[1] (or antisymmetric ) matrix and is the term that represents the magnetic field. This also shows that MHMC only differs from HMC dynamics in equation (4) by $\mathbf{G}$ being non-zero. When $\mathbf{G} = \mathbf{0}$, MHMC and HMC have the same dynamics. Tripuraneni *et al.* [4] prove that in three dimensions, these dynamics are Newtonian mechanics of a charged particle in a magnetic field. How this magnetic field relates to the force field (e.g. are they orthogonal?) will determine the extent of the sampling efficiency of MHMC over HMC [24].

As with HMC, these non-canonical dynamics cannot be integrated exactly, and we resort to a numerical integration scheme with a MH acceptance step to ensure detailed balance. The update equations for the leapfrog-like integration scheme for MHMC, for the case where $\mathbf{M} = \mathbf{I}$, are given as [4]:

$$\begin{aligned} \mathbf{p}_{t+\frac{\epsilon}{2}} &= \mathbf{p}_t + \frac{\epsilon}{2}\frac{\partial H\left(\mathbf{w}_t, \mathbf{p}_t\right)}{\partial \mathbf{w}} \\ \mathbf{w}_{t+\epsilon} &= \mathbf{w}_t + \mathbf{G}^{-1}\left(\exp\left(\mathbf{G}\epsilon\right) - \mathbf{I}\right)\mathbf{p}_{t+\frac{\epsilon}{2}} \\ \mathbf{p}_{t+\frac{\epsilon}{2}} &= \exp\left(\mathbf{G}\epsilon\right)\mathbf{p}_{t+\frac{\epsilon}{2}} \\ \mathbf{p}_{t+\epsilon} &= \mathbf{p}_{t+\frac{\epsilon}{2}} + \frac{\epsilon}{2}\frac{\partial H\left(\mathbf{w}_{t+\epsilon}, \mathbf{p}_{t+\frac{\epsilon}{2}}\right)}{\partial \mathbf{w}}. \end{aligned} \tag{10}$$

The above equations show that we can retrieve the update equations of traditional HMC by first performing a Taylor matrix expansion for the exponential and then substituting $\mathbf{G} = 0$. The pseudo-code for the MHMC algorithm is shown in Algorithm 2. It is important to note that we need to flip the sign of $\mathbf{G}$ (see lines 8-15 in Algorithm 2), as we do the sign of $\mathbf{p}$ in HMC, so as to render the MHMC algorithm reversible. In this sense, we treat $\mathbf{G}$ as being an auxiliary variable in the same fashion as $\mathbf{p}$ [4]. In this setup, $\mathbf{p}$ would be Gaussian while $\mathbf{G}$ would have a binary distribution [4] and only taking on the values $\pm\mathbf{G_0}$, with $\mathbf{G_0}$ being specified by the user. Exploring more complex distributions for $\mathbf{G}$ is still an open area of research.

Although MHMC requires matrix exponentiation and inversion as shown in equation (10), this only needs to be computed once upfront and stored [4]. Following this approach results in computation time that is comparable to HMC, which becomes more important in models that have many parameters such as neural networks.

As $\mathbf{G}$ only needs to be antisymmetric, there is no guarantee that it will be invertible. In this case, we need first to diagonalise $\mathbf{G}$ and separate its invertible or singular components [4]. As $\mathbf{G}$ is strictly antisymmetric, we can express it as $i\mathbf{H}$ where $\mathbf{H}$ is a Hermitian matrix, and can thus be diagonilised over the space of complex numbers $\mathbb{C}$ as [4]:

$$\mathbf{G} = \left[W_{\mathbf{\Lambda}} \quad W_{\mathbf{0}}\right]\begin{bmatrix}\mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}W_{\mathbf{\Lambda}}^T \\ W_{\mathbf{0}}^T\end{bmatrix} \tag{11}$$

where $\mathbf{\Lambda}$ is a diagonal submatrix consisting of the nonzero eigenvalues of $\mathbf{G}$, columns of $W_{\mathbf{\Lambda}}$, and $W_{\mathbf{0}}$ are the eigenvectors of $\mathbf{G}$ corresponding to its nonzero and zero eigenvalues,

[1] That is: $\mathbf{G}^T = -\mathbf{G}$.

respectively. This leads to the following update for $\mathbf{w}$ in equation (10) [4]:

$$\mathbf{w}_{t+\epsilon} = \mathbf{w}_t + [W_{\mathbf{\Lambda}} \quad W_{\mathbf{0}}] \times \begin{bmatrix} \mathbf{\Lambda}^{-1}\left(\exp(\mathbf{\Lambda}\epsilon) - \mathbf{I}\right) & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} W_{\mathbf{\Lambda}}^T \\ W_{\mathbf{0}}^T \end{bmatrix} \mathbf{p}_{t+\frac{\epsilon}{2}} \qquad (12)$$

It is worthwhile noting that when $\mathbf{G} = 0$ in equation (12) then the flow map will reduce to an Euler translation as in traditional HMC [4].

**Algorithm 2:** Magnetic Hamiltonian Monte Carlo Algorithm

```
    Input: $N, \epsilon, L, w_{\text{init}}, H(w, p), G$
    Output: $(w)_{m=0}^{N}$
 1: $w_0 \leftarrow w_{\text{init}}$
 2: for $m \to 1$ to $N$ do
 3:     $p_{m-1} \sim \mathcal{N}(0, \mathbf{M})$
 4:     $p_m, w_m =$ Integrator$(p_{m-1}, w_{m-1}, \epsilon, L, G, H)$ in
        equation (10)
 5:     $\delta H = H(w_{m-1}, p_{m-1}) - H(w_m, p_m)$
 6:     $\alpha_m = \min(1, \exp(\delta H))$
 7:     $u_m \sim \text{Unif}(0, 1)$
 8:     if $\alpha_m > u_m$ then
 9:         $w_m = w_m$
10:         $G = -G, p_m = -p_m$
11:     else
12:         $w_m = w_{m-1}$
13:     end if
14:     $p_m = -p_m \leftarrow$ flip momentum
15:     $G = -G \leftarrow$ flip magnetic field
16: end for
```

## III. SHADOW HAMILTONIAN FOR MAGNETIC DYNAMICS

The leapfrog-like integrator for MHMC only preserves the Hamiltonian, that is, the total energy of the system, up to second order $\mathcal{O}(\epsilon^2)$ [18], [18], [19]. This leads to a larger than expected value for $\delta H$ in line 5 of Algorithm 2 for long trajectories, which results in more rejections in the MH step in line 8 of Algorithm 2. To increase the accuracy of the preservation of the total energy to higher orders, and consequently maintain high acceptance rates, one could: 1) decrease the step size and thus only consider short trajectories, or 2) utilise numerical integration schemes which preserve the Hamiltonian to a higher order, 3) or a combination of 1) and 2). These three approaches typically lead to a high computational burden, which is not ideal [5], [23].

An alternative strategy is to assess the error produced by feeding the solution backward through [21], [27] the leapfrog-like integration scheme in equation (10), to derive a modified Hamiltonian whose energy is preserved to a higher-order by the integration scheme than the true Hamiltonian [21]. This modified Hamiltonian is also referred to as the shadow Hamiltonian. We then sample from the shadow density and correct for the induced bias via importance sampling as is done in [5], [18], [19], [23], among others.

Shadow Hamiltonians are perturbations of the Hamiltonian that are by design exactly conserved by the numerical integrator [19]–[21], [23]. In this manuscript, we focus on a fourth-order truncation of the shadow Hamiltonian under the leapfrog-like numerical integrator in equation (10). Since the MHMC numerical integrator is second-order accurate ($\mathcal{O}^2$) [4], the fourth-order truncation is conserved with higher accuracy ($\mathcal{O}^4$) by the integrator than the true Hamiltonian. In Theorem 1, we derive the fourth-order shadow Hamiltonian under the numerical integrator.

*Theorem 1: Let $H : R^d \times R^d = R$ be a smooth Hamiltonian function. The fourth–order shadow Hamiltonian function $\hat{H} : R^d \times R^d = R$ corresponding to the numerical integrator used in MHMC is given by:*

$$\hat{H}(\mathbf{w}, \mathbf{p}) = H(\mathbf{w}, \mathbf{p}) + \frac{\epsilon^2}{12}\left[K_{\mathbf{p}} U_{\mathbf{ww}} K_{\mathbf{p}} + K_{\mathbf{p}} \mathbf{G} K_{\mathbf{pp}} U_{\mathbf{w}}\right] - \frac{\epsilon^2}{24}\left[U_{\mathbf{w}} K_{\mathbf{pp}} U_{\mathbf{w}}\right] + \mathcal{O}(\epsilon^4) \qquad (13)$$

*Proof:* As outlined in Tripuraneni *et al.* [4], the Hamiltonian vector field $\overrightarrow{H} = \nabla_{\mathbf{p}} H \nabla_{\mathbf{w}} + (-\nabla_{\mathbf{w}} + \mathbf{G}\nabla_{\mathbf{p}} H)\nabla_{\mathbf{p}} = \overrightarrow{A} + \overrightarrow{B}$ will generate the exact flow corresponding to exactly simulating the MHMC dynamics [4]. We obtain the shadow Hamiltonian via the separability of the true Hamiltonian [4]. The numerical integration scheme in equation (10) splits the Hamiltonian as: $H(\mathbf{w}, \mathbf{p}) = H_1(\mathbf{w}) + H_2(\mathbf{p}) + H_1(\mathbf{w})$ and exactly integrates each sub-Hamiltonian [4]. Using the Baker-Campbell-Hausdorff [36] formula we obtain:

$$\begin{aligned} \Phi_{\epsilon,H}^{frog} &= \Phi_{\epsilon,H_1(\mathbf{w})} \circ \Phi_{\epsilon,H_2(\mathbf{p})} \circ \Phi_{\epsilon,H_1(\mathbf{w})} \\ &= \exp\left(\frac{\epsilon}{2}\overrightarrow{B}\right) \circ \exp\left(\epsilon\overrightarrow{A}\right) \circ \exp\left(\frac{\epsilon}{2}\overrightarrow{B}\right) \\ &= H(\mathbf{w}, \mathbf{p}) + \frac{\epsilon^2}{12}\{K, \{K, U\}\} \\ &\quad - \frac{\epsilon^2}{24}\{U, \{U, K\}\} + \mathcal{O}(\epsilon^4) \end{aligned} \qquad (14)$$

where the non-canonical Poisson brackets [6], [37] are defined as:

$$\begin{aligned} \{f, g\} &= [\nabla_{\mathbf{w}} f, \nabla_{\mathbf{p}} f] \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{G} \end{bmatrix} [\nabla_{\mathbf{w}} g, \nabla_{\mathbf{p}} g]^T \\ &= -\nabla_{\mathbf{p}} f \nabla_{\mathbf{w}} g + \nabla_{\mathbf{w}} f \nabla_{\mathbf{p}} g + \nabla_{\mathbf{p}} f \mathbf{G} \nabla_{\mathbf{p}} g \end{aligned} \qquad (15)$$

and collapse to the canonical Poisson brackets when $\mathbf{G} = \mathbf{0}$ [6], [37]. The corresponding derivatives from the non-canonical Poisson brackets are presented in Appendix A. The shadow Hamiltonian for the leapfrog-like integrator is then:

$$\hat{H}(\mathbf{w}, \mathbf{p}) = H(\mathbf{w}, \mathbf{p}) + \frac{\epsilon^2}{12}\left[K_{\mathbf{p}} U_{\mathbf{ww}} K_{\mathbf{p}} + \underbrace{K_{\mathbf{p}} \mathbf{G} K_{\mathbf{pp}} U_{\mathbf{w}}}_{A}\right] - \frac{\epsilon^2}{24}\left[U_{\mathbf{w}} K_{\mathbf{pp}} U_{\mathbf{w}}\right] + \mathcal{O}(\epsilon^4) \qquad (16)$$

where $A$ is the factor induced by the presence of the magnetic field $\mathbf{G}$. When $\mathbf{G} = \mathbf{0}$, the shadow in (16) becomes the shadow for Hamiltonian dynamics [18], [19]. Note that the modified Hamiltonian in (16) is preserved to fourth-order [5], [18], [23], and is thus more accurately conserved by numerical integrator in MHMC than the true Hamiltonian [38]. □

## IV. SHADOW MAGNETIC HAMILTONIAN MONTE CARLO

We now present the Shadow Magnetic Hamiltonian Monte Carlo (SMHMC) algorithm, which combines non-canonical Hamiltonian dynamics in MHMC with the high conservation property of shadow Hamiltonians. The benefits of employing non-canonical Hamiltonian dynamics in MHMC have already been established in [4], [17], [24], while the advantages of shadow Hamiltonians in general are presented in [5], [18], [20]–[22]. We combine these two concepts to create a new sampler that outperforms MHMC across various performance metrics.

An analysis of the shadow Hamiltonian corresponding to MHMC in equation (13) shows that the conditional density for the momenta $\pi_H(\mathbf{p}|\mathbf{w})$ is not Gaussian. This suggests that if we fully re-sample the momenta from a normal distribution, as is done in [33], we will attain a sampler that does not satisfy detailed balance [5], [22]. This necessitates computationally intensive momentum generation [19] or partial momentum refreshment [22], [23] with an MH step. In this manuscript, we utilise the partial momentum refreshment procedure outlined in [5], [22], in which a Gaussian noise vector $u \sim \mathcal{N}(0, \mathbf{M})$ is drawn. The momentum proposal is then produced via the mapping:

$$R(\mathbf{p}, u) = \left(\rho\mathbf{p} + \sqrt{1-\rho^2}u, -\sqrt{1-\rho^2}\mathbf{p} + \rho u\right) \tag{17}$$

The new parameter, which we refer to as the momentum refreshment parameter, $\rho = \rho(\mathbf{w}, \mathbf{p}, u)$ takes values between zero and one, and controls the extent of the momentum retention [5], [17], [23]. When $\rho$ is equal to one, the momentum is never updated and when $\rho$ is equal to zero, the momentum is always updated [17]. The momentum proposals are then accepted according to the modified non-separable shadow Hamiltonian given as $\bar{H}(\mathbf{w}, \mathbf{p}, u) = \hat{H}(\mathbf{w}, \mathbf{p}) + \frac{1}{2}u\mathbf{M}^{-1}u$. The updated momentum is then taken to be $\rho\mathbf{p} + \sqrt{1-\rho^2}u$ with probability:

$$\omega := \max\{1, \exp(\bar{H}(\mathbf{w}, \mathbf{p}, u) - \bar{H}(\mathbf{w}, R(\mathbf{p}, u)))\}. \tag{18}$$

The incomplete refreshment of the momentum produces a chain which saves some of the behaviour between neighbourhood samples [5], [17], [22], [23], [39]. In Section V-D, we assess the sensitivity of the sampling results on the user-specified value of $\rho$. An algorithmic description of the SMHMC sampler is provided in Algorithm 3. It is worth noting from Algorithm 3 that the SMHMC sampler uses two reversible MH steps, which implies that the resulting Markov chain is no longer reversible [5], [22], [23]. By breaking the detailed balance condition, it is no longer immediately clear that the target density is stationary, and so this must be demonstrated [5], [23].

*Theorem 2: The SMHMC algorithm leaves the importance target distribution invariant.*

*Proof:* The proof of theorem 2 is obtained in Appendix A of the paper by Radivojevic and Akhmatskay [23]. The proof involves showing that the addition of step 4 in Algorithm 3 leaves the target invariant. The result follows from [23] by making use of the fact that the explicit form of the shadow Hamiltonian, which has the additional magnetic component $A = K_{\mathbf{p}}\mathbf{G}K_{\mathbf{pp}}U_{\mathbf{w}}$ in equation (16) in our case, is not required for the proof [5], [23]. □

**Algorithm 3:** Shadow Magnetic Hamiltonian Monte Carlo Algorithm

```
    Input: $L$, $\epsilon$, $\rho$, $N$, $\mathbf{G}$, $(\mathbf{w_0}, \mathbf{p_0})$.
    Output: $(w_i, p_i, b_i)_{i=0}^{N}$
 1: for $i \to 1$ to $N$ do
 2:   $(\mathbf{w}, \mathbf{p}) \leftarrow (\mathbf{w_{i-1}}, \mathbf{p_{i-1}})$.
 3:   $u \sim \mathcal{N}(0, \mathbf{M})$
 4:   $\bar{p} \leftarrow \rho p + \sqrt{1-\rho^2}u$ with probability $\omega$ in equation
      (18).
 5:   $(\hat{\mathbf{w}}, \hat{\mathbf{p}}) = \Phi^{L}_{\epsilon,H}(\mathbf{w}, \bar{\mathbf{p}}, \mathbf{G})$ in equation (10)
 6:   $\zeta = \min\left[1, \exp(-\delta\hat{H})\right]$, $u \sim \text{Unif}(0, 1)$
 7:   if $\zeta > u$ then
 8:     $(\mathbf{w}_i, \mathbf{p}_i, \mathbf{G}) \leftarrow (\hat{\mathbf{w}}, \hat{\mathbf{p}}, \mathbf{G})$
 9:   else
10:     $(\mathbf{w}_i, \mathbf{p}_i, \mathbf{G}) \leftarrow (\mathbf{w}, -\mathbf{p}, -\mathbf{G})$
11:   end if
12:   $b_i = \exp\left(\hat{H}(\mathbf{w_i}, \mathbf{p_i}) - H(\mathbf{w_i}, \mathbf{p_i})\right)$
13: end for
```

## V. EXPERIMENT DESCRIPTION

In this section, we outline the settings used for the experiments, the performance metrics, approach for algorithm tuning, and we also present the sensitivity analysis for the partial momentum refreshment parameter $\rho$ used in SMHMC and PMHMC.

### A. EXPERIMENT SETTINGS

In our analysis, we compare the performance of SMHMC against MHMC and PMHMC across a multivariate Gaussian distribution described in [17] with $D = 10$, the Protein dataset [20] modeled using a Bayesian Neural Network (BNN) and the Heart [20] and Pima [20] datasets modeled using Bayesian Logistic Regression (BLR). The details of the real-world datasets are shown in Table 1. Note that the BNN architecture used is an MLP with one hidden layer and five hidden units. For all the target posteriors used in this paper, the momentum refreshment parameter $\rho$ is set to 0.7. This setting worked well on all the targets. Further experiments of the sensitivity to $\rho$ are presented in Section V-D.

### B. PERFORMANCE METRICS

We now present the performance metrics used to measure the performance of the algorithms proposed in this manuscript.

The performance metrics used are the acceptance rate, the multivariate effective sample size (ESS), the multivariate ESS normalised by the execution time. We also assess the convergence of the proposed SMHMC method using the potential scale reduction factor metric. The acceptance rate metric measures the number of generated samples that are accepted in the MH acceptance step of the algorithm [23], [29]. The higher the number of accepted samples for the same step size, the more preferable the method. We discuss the remaining metrics in more detail in the following sections.

#### 1) EFFECTIVE SAMPLE SIZE

The ESS metric is a commonly used metric for assessing the sampling efficiency of an MCMC algorithm. It indicates the number of effectively uncorrelated samples out of the total number of generated samples [23], [29]. The larger the ESS, the better the performance of the MCMC method. The ESS normalised by execution time metric takes into account the computational resources required to generate the samples and penalises MCMC methods that require more computational resources to generate the same number of uncorrelated samples. The larger this metric, the better the efficiency of the algorithm.

This paper employs the multivariate ESS metric developed by Vats *et al.* [40] instead of the minimum univariate ESS metric typically used in analysing MCMC results. The minimum univariate ESS measure is not able to capture the correlations between the different parameter dimensions, while the multivariate ESS metric can incorporate this information [2], [20], [40]. The minimum univariate ESS calculation results in the estimate of the ESS being dominated by the parameter dimensions that mix the slowest and ignore all other dimensions [20], [40]. The multivariate ESS is calculated as:

$$\text{mESS} = N \times \left(\frac{|\Lambda|}{|\Sigma|}\right)^{\frac{1}{D}} \tag{19}$$

where $N$ is the number of generated samples, $D$ is the number of parameters, $|\Lambda|$ is the determinant of the sample covariance matrix and $|\Sigma|$ is the determinant of the estimate of the Markov chain standard error. When $D = 1$, mESS is equivalent to the univariate ESS measure [40]. Note that when there are no correlations in the chain, we have that $|\Lambda| = |\Sigma|$ and mESS $= N$.

We now address the ESS calculation for Markov chains that have been re-weighted via importance sampling, such is the case for the SMHMC algorithm proposed in this paper [5], [20], [21], [23]. For $N$ samples re-weighted by importance sampling, the common approach is to use the approximation by Kish [5], [41] given by

$$\text{ESS}_{IMP} = \frac{1}{\left(\sum_{j=1}^{N} \bar{b}_j^2\right)} \tag{20}$$

where $\bar{b}_j = b_j / \sum_{k=1}^{N} b_k$. This accounts for the possible non-uniformity in the importance sampling weights. In order to account for both the effects of sample auto-correlation and re-weighting via importance sampling, we approximate ESS under importance sampling by taking directions from Heide *et al.* [5] and using:

$$\text{ESS} := \frac{\text{ESS}_{IMP}}{N} \times \text{mESS} = \frac{1}{\left(\sum_{j=1}^{N} \bar{b}_j^2\right)} \times \left(\frac{|\Lambda|}{|\Sigma|}\right)^{\frac{1}{D}} \tag{21}$$

#### 2) CONVERGENCE ANALYSIS

The $\hat{R}$ diagnostic of Gelman and Rubin [42] is a favored technique for verifying the convergence of MCMC chains [43]. This diagnostic depends on running numerous chains $\{X_{i0}, X_{i1}, \ldots, X_{i(N-1)}\}$ for $i \in \{1, 2, 3, \ldots, m\}$ starting at diverse initial conditions with $m$ being the number of chains and $N$ being the sample size. Using these parallel chains, two estimators of the variance can be constructed. The estimators are the between-the-chain variance estimate and the within-the-chain variance. When the chain has converged, the ratio of these two estimators should be one. The $\hat{R}$ metric, which is formally known as the potential scale reduction factor, is defined as:

$$\hat{R} = \frac{\hat{V}}{W} \tag{22}$$

where

$$W = \sum_{i=1}^{m} \sum_{j=0}^{N-1} \frac{\left(X_{ij} - \bar{X}_{i.}\right)^2}{m(N-1)} \tag{23}$$

is the within-chain variance estimate and $\hat{V} = \frac{N-1}{N} W + \frac{B}{N}$ is the pooled variance estimate which incorporates the between-chains

$$B = \sum_{j=0}^{N-1} \frac{\left(\bar{X}_{i.} - \bar{X}_{..}\right)^2}{m-1} \tag{24}$$

and within-chain $W$ variance estimates, with $\bar{X}_{i.}$ and $\bar{X}_{..}$ being the $i^{th}$ chain mean and overall mean respectively for $i \in \{1, 2, 3, \ldots, m\}$. Values larger than the convergence threshold of 1.05 for the $\hat{R}$ metric indicate divergence of the chain [6], [42]. In this paper, we asses the convergence of the chains by computing the *maximum* $\hat{R}$ metric over each of the parameter dimensions for the given target.

### C. ALGORITHM PARAMETER TUNING

As mentioned in Section II, the matrix $\mathbf{G}$ in the MHMC method provides an extra degree of freedom which typically results in better sampling behavior than HMC [4], [20], [25]. It is not immediately clear how this matrix should be set - this is still an open area of research [4], [6], [20]. In this paper, we take direction from the inventors [4] of the method and select only a few dimensions to be influenced by the magnetic field. In particular, $\mathbf{G}$ was set such that $\mathbf{G_{1i}} = g$, $\mathbf{G_{i1}} = -g$ and zero elsewhere where $g = 0.2$ for the BLR datasets, and $g = 0.1$ for all the other targets.

These settings mean that the selection of $\mathbf{G}$ is not necessarily the optimal choice for all the target distributions

**TABLE 1.** Real-world datasets used in this paper. *N* represents the number of observations. BLR is Bayesian Logistic Regression, and BNN means Bayesian Neural Networks. *D* denotes the number of model parameters.

| Dataset | Features | $N$ | Model | $D$ |
|---|---|---|---|---|
| Pima | 7 | 532 | BLR | 8 |
| Heart | 13 | 270 | BLR | 14 |
| Australian | 14 | 690 | BLR | 15 |
| Protein | 9 | 45 730 | BNN | 56 |

**TABLE 2.** Step size and trajectory length parameters used for the MCMC methods in this manuscript. Five thousand samples were used to tune the step size for the given trajectory length using primal-dual averaging. The target acceptance rate was set to 80%.

| **Problem** | $L$ | $\epsilon$ |
|---|---|---|
| Gaussian with $D = 10$ | 15 | 0.1414 |
| Pima | 50 | 0.0300 |
| Heart | 50 | 0.0241 |
| Australian | 50 | 0.0300 |
| Protein | 25 | 0.0691 |

considered but was adequate for our objectives as this basic setting still leads to satisfactory performance on the SMHMC algorithm that we present in this manuscript. Tuning **G** for each target posterior should result in improved performance compared to the results given in this manuscript. An alternative approach to the selection of **G** would have been to follow [6] and selecting **G** to be a random antisymmetric matrix. It is not immediately clear if the approach of [6] is necessary optimal, and we plan to explore this approach in future work.

The step size $\epsilon$ is tuned to target an acceptance rate of 80% using the primal-dual averaging methodology of [3]. The trajectory lengths $L$ used vary across the different targets, with the final step sizes and trajectory lengths used for the diverse problems presented in Table 2.

Ten independent chains were run for each approach on each target distribution. Three thousand samples were generated for each target, with the first one-thousand samples discarded as burn-in. These settings were adequate for all the methods to converge on all the target posteriors. All the experiments in this manuscript were conducted on a machine with a 64bit CPU using PyTorch.

### D. SENSITIVITY TO MOMENTUM REFRESHMENT PARAMETER

We investigate the effects of varying the momentum refreshment parameter $\rho$ on the sampling performance of the proposed shadow Hamiltonian method. Ten chains, starting from different positions, of the PMHMC and shadow MHMC algorithms were ran on the Australian credit dataset for $\rho \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. Note that we exclude $\rho = 1.0$ as the chain is likely to no longer be ergodic as the momentum would not be refreshed at all [5]. Figure 2 shows the results for the Australian credit dataset. The results show that PMHMC and SMHMC have stable acceptance rates across the different values of $\rho$ on all this target. SMHMC has higher acceptance rates and ESS than PMHMC for the same step size. However, due to the high execution time of SMHMC, PMHMC produced better time-normalised ESS compared to SMHMC. The methods show a general trend of increasing ESS and time-normalised ESS with increasing $\rho$.

**TABLE 3.** Multivariate Gaussian distribution with *D* = 10 results averaged over ten runs. The time *t* is in seconds. The values in **bold** indicate that the particular method outperforms the other methods on that specific metric. AR stands for the acceptance rate of the generated samples post the burn-in period.

| Gaussian with $D = 10$ | | | |
|---|---|---|---|
| Metric | MHMC | PMHMC | SMHMC |
| AR | 80.96 | 82.86 | **84.64** |
| ESS | 1 844 | 1 936 | **2 546** |
| $t$ | **19.52** | **19.52** | 69.01 |
| ESS/$t$ | 94.45 | **99.17** | 36.90 |
| $\hat{R}$ max | 1.02 | 1.02 | **1.01** |

**TABLE 4.** Protein dataset results averaged over ten runs. The time *t* is in seconds. The values in **bold** indicate that the particular method outperforms the other methods on that specific metric. AR stands for the acceptance rate of the generated samples post the burn-in period.

| Protein Dataset | | | |
|---|---|---|---|
| Metric | MHMC | PMHMC | SMHMC |
| AR | 80.28 | 79.4 | **82.09** |
| ESS | 1 729 | 2 467 | **3 244** |
| $t$ | **373** | 374 | 1 579 |
| ESS/$t$ | 4.51 | **6.54** | 2.05 |
| $\hat{R}$ max | 1.01 | **1.00** | **1.00** |

**TABLE 5.** Heart dataset results averaged over ten runs. The time *t* is in seconds. The values in **bold** indicate that the particular method outperforms the other methods on that specific metric. AR stands for the acceptance rate of the generated samples post the burn-in period.

| Heart Dataset | | | |
|---|---|---|---|
| Metric | MHMC | PMHMC | SMHMC |
| AR | 79.10 | 83.67 | **84.66** |
| ESS | 2 864 | 4 215 | **4 350** |
| $t$ | 68.24 | **67.5** | 71.90 |
| ESS/$t$ | 41.96 | **62.42** | 60.50 |
| $\hat{R}$ max | 1.01 | 1.01 | **1.00** |

## VI. RESULTS AND DISCUSSION

Figure 3 shows the diagnostic trace-plots of the negative log-likelihood across various target posteriors. The results show that all three methods have converged on the four target densities analysed in this paper.

The performance of the algorithms across different metrics is shown in Figure 4 and Tables 3 to 6. In Figure 4, the plots on the first row for each dataset show the effective sample size, and the plots on the second row show the effective sample size normalised by execution time. The results are for the ten runs of each algorithm. The execution time $t$ in Figure 4 and Tables 3 to 6 is in seconds. The results in Tables 3 to 6 are the mean results over the ten runs for each algorithm.

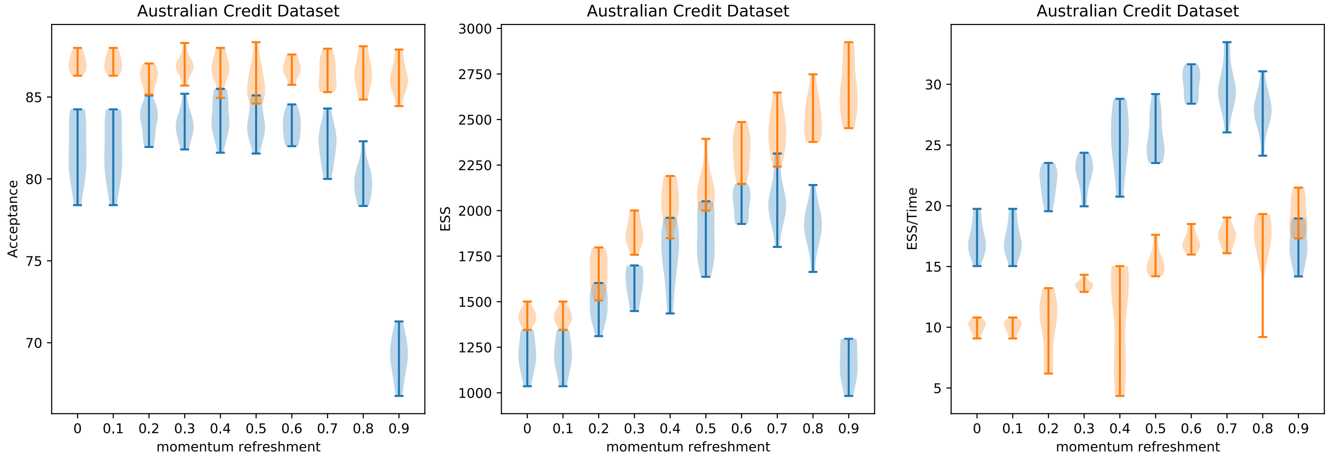

**FIGURE 2.** Acceptance rates, ESS and ESS/Time for ten chains of PMHMC (blue) and SMHMC (orange) the Australian credit dataset with varying choices of $\rho$. The ESS metrics are an increasing function of $\rho$ with the acceptance rate of SMHMC being larger than PMHMC for the same step size $\epsilon$.

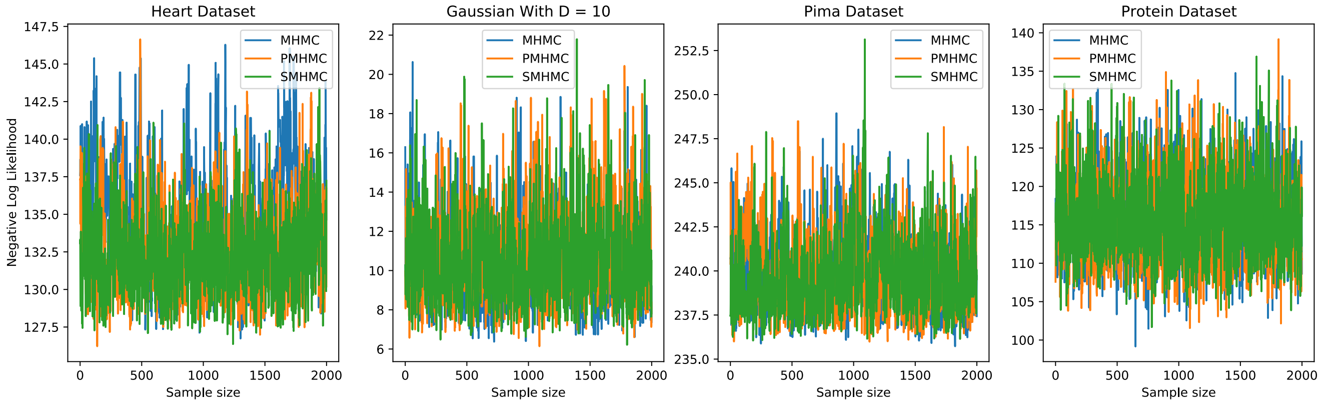

**FIGURE 3.** Diagnostic trace-plots of the negative log-likelihood across various targets averaged over ten runs of each method. These results show that all the MCMC methods have converged on all the targets.

**TABLE 6.** Pima dataset results averaged over ten runs. The time $t$ is in seconds. The values in bold indicate that the particular method outperforms the other methods on that specific metric. AR stands for the acceptance rate of the generated samples post the burn-in period.

| Pima Dataset | | | |
|---|---|---|---|
| Metric | MHMC | PMHMC | SMHMC |
| AR | 82.35 | 80.60 | **84.80** |
| ESS | 902 | 1 595 | **1 699** |
| $t$ | 84.13 | **72.91** | 110.66 |
| ESS/$t$ | 10.72 | **21.88** | 15.35 |
| $\hat{R}$ max | 1.02 | **1.01** | **1.01** |

We use the mean values over the ten runs in Tables 3 to 6 to form our conclusions about the performance of the algorithms.

Tables 3 to 6 show that SMHMC produces the highest acceptance rate across all the targets. Furthermore, the SMHMC algorithm produces the largest ESS on all the targets. In particular, it outperforms PMHMC, which shows that the method is doing something extra than just incorporating partial momentum refreshment into the MHMC. However, the source of the outperformance of SMHMC seems to be mostly driven by the incorporation of the partial momentum refreshment, highlighting the benefits of utilising partial momentum refreshment in Hamiltonian dynamics-based samplers in general.

The results show that the MHMC and PMHMC produce the lowest execution times across all the targets, with SMHMC having the largest execution time, sometimes as much as two times that of the MHMC method. The large execution time of SMHMC can be attributed to the multiple times that the shadow Hamiltonian is evaluated, as well as the extra MH step for the momenta generation. The slow execution time is the key drawback of SMHMC, and hinders the performance of the method on a time-normalised ESS basis. We find that PMHMC outperforms all the methods on a time-normalised ESS basis. SMHMC outperforms MHMC on the BLR datasets on a time-normalised ESS basis, with MHMC outperforming SMHMC on the other targets on the same basis. Furthermore, the $\hat{R}$ metric shows that all

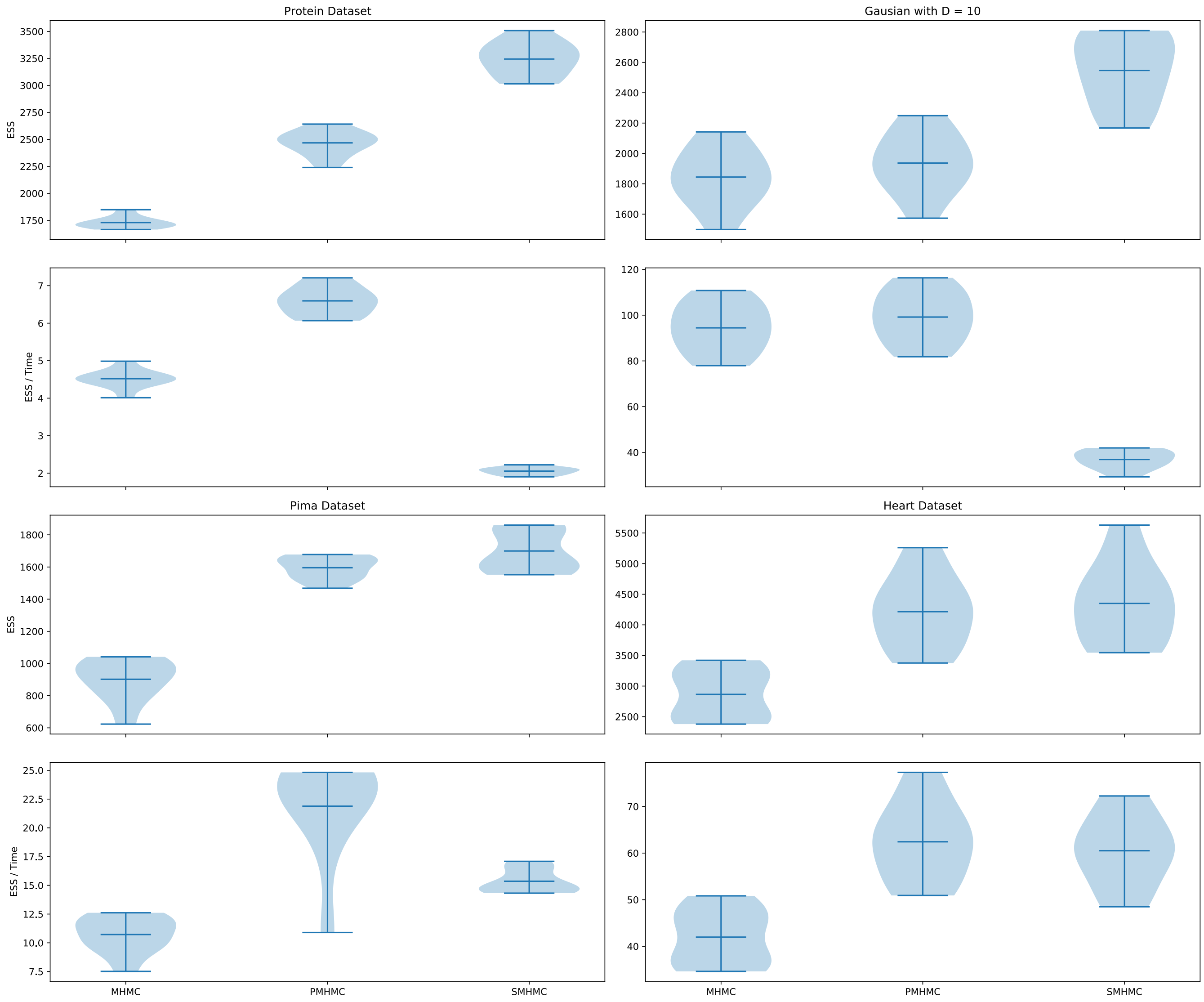

**FIGURE 4.** Results for the datasets over ten runs of each method. For each dataset, the plots on the first row show the multivariate effective sample size and the plots on the second row show the multivariate effective sample size normalised by execution time (in seconds). For all the plots, the larger the value, the better the method. The dark horizontal line in each violin plot represents the mean value over ten runs of each algorithm.

the methods have converged, with PMHMC and SMHMC producing marginally better convergence behaviour compared to MHMC.

## VII. CONCLUSION

We introduce the novel SMHMC algorithm, which combines the non-canonical dynamics of MHMC with the benefits of sampling from a shadow Hamiltonian. This combination results in improved exploration of the posterior when compared to MHMC. The empirical results show that the new algorithm provides a significant improvement on the MHMC algorithm in terms of higher acceptance rates and larger effective sample sizes, even as the dimensionality of the problem increases.

A primary limitation of the proposed algorithm is the computational time associated with the method, mainly since it involves the computation of the Hessian matrix of the target distribution. This leads to poor performance on a time-normalised ESS basis. A straightforward approach to circumvent the computational burden is to use closed-form expressions for the first-order derivatives and the Hessian matrix. This approach, however, restricts the possible targets that can be considered. We aim to address this issue in the future by using a surrogate model to approximate the shadow Hamiltonian during the burn-in period of the method as an active learning task.

Another limitation of the method is the need to tune the momentum refreshment parameter. Although typically higher values of the parameter improve the effective sample sizes, a more robust approach to selecting the parameter is still required. In future work, we plan on improving the proposed method by establishing an automated process to tune the momentum refreshment parameter.

## APPENDIX

### A. DERIVATIVES FROM POISSON BRACKETS

We now present the derivatives derived from the non-canonical Poisson brackets:

$$\begin{aligned}\{K, U\} &= -\nabla_{\mathbf{p}} K \nabla_{\mathbf{w}} U + \nabla_{\mathbf{w}} K \nabla_{\mathbf{p}} U \\ &\quad + \nabla_{\mathbf{p}} K \mathbf{G} \nabla_{\mathbf{p}} U - K_{\mathbf{p}} U_{\mathbf{w}} \\ \{K, \{K, U\}\} &= -K_{\mathbf{p}} U_{\mathbf{ww}} K_{\mathbf{p}} - K_{\mathbf{p}} \mathbf{G} K_{\mathbf{pp}} U_{\mathbf{w}} \\ \{U, K\} &= U_{\mathbf{w}} K_{\mathbf{p}} \\ \{U, \{U, K\}\} &= U_{\mathbf{w}} K_{\mathbf{pp}} U_{\mathbf{w}} \qquad (25)\end{aligned}$$

## ACKNOWLEDGMENT

The computations in this work were performed on resources provided by the Center for High Performance Computing (CHPC) at the Council of Scientific and Industrial Research (CSIR) South Africa.

## REFERENCES

[1] R. M. Neal, *Bayesian Learning for Neural Networks*, vol. 118. New York, NY, USA: Springer-Verlag, 2012.

[2] M. Girolami and B. Calderhead, "Riemann manifold Langevin and Hamiltonian Monte Carlo methods," *J. Roy. Statist. Soc. B, Stat. Methodol.*, vol. 73, no. 2, pp. 123–214, Mar. 2011.

[3] M. D. Hoffman and A. Gelman, "The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1593–1623, Apr. 2014.

[4] N. Tripuraneni, M. Rowland, Z. Ghahramani, and R. Turner, "Magnetic Hamiltonian Monte Carlo," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3453–3461.

[5] C. Heide, F. Roosta, L. Hodgkinson, and D. Kroese, "Shadow manifold Hamiltonian Monte Carlo," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 1477–1485.

[6] J. A. Brofos and R. R. Lederman, "Non-canonical Hamiltonian Monte Carlo," 2020, *arXiv:2008.08191*.

[7] G. E. Hinton and D. van Camp, "Keeping the neural networks simple by minimizing the description length of the weights," in *Proc. 6th Annu. Conf. Comput. Learn. Theory (COLT)*, 1993, pp. 5–13.

[8] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *J. Mach. Learn. Res.*, vol. 14, no. 5, pp. 1303–1347, 2013.

[9] F. Ruiz and M. Titsias, "A contrastive divergence for combining variational inference and MCMC," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5537–5545.

[10] T. Salimans, D. Kingma, and M. Welling, "Markov chain Monte Carlo and variational inference: Bridging the gap," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1218–1226.

[11] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, Apr. 1970.

[12] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, "Hybrid Monte Carlo," *Phys. Lett. B*, vol. 195, pp. 216–222, Sep. 1987.

[13] M. Graham and A. Storkey, "Continuously tempered Hamiltonian Monte Carlo," in *Proc. Conf. Uncertainty Artif. Intell. (UAI)*, 2017. [Online]. Available: http://auai.org/uai2017/proceedings/papers/289.pdf

[14] Z. Liu and Z. Zhang, "Quantum-inspired Hamiltonian Monte Carlo for Bayesian sampling," 2019, *arXiv:1912.01937*.

[15] M. Graham and A. Storkey, "Continuously tempered Hamiltonian Monte Carlo," in *Proc. Adv. Approx. Bayesian Inference, NIPS Workshop*, 2016. [Online]. Available: http://approximateinference.org/accepted/GrahamStorkey2016.pdf

[16] R. Luo, J. Wang, Y. Yang, J. Wang, and Z. Zhu, "Thermostat-assisted continuously-tempered Hamiltonian Monte Carlo for Bayesian learning," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Red Hook, NY, USA: Curran, 2018, pp. 10673–10682.

[17] W. T. Mongwe, R. Mbuvha, and T. Marwala, "Magnetic Hamiltonian Monte Carlo with partial momentum refreshment," *IEEE Access*, vol. 9, pp. 108009–108016, 2021.

[18] C. R. Sweet, S. S. Hampton, R. D. Skeel, and J. A. Izaguirre, "A separable shadow Hamiltonian hybrid Monte Carlo method," *J. Chem. Phys.*, vol. 131, no. 17, Nov. 2009, Art. no. 174106.

[19] J. A. Izaguirre and S. S. Hampton, "Shadow hybrid Monte Carlo: An efficient propagator in phase space of macromolecules," *J. Comput. Phys.*, vol. 200, no. 2, pp. 581–604, Nov. 2004.

[20] W. T. Mongwe, R. Mbuvha, and T. Marwala, "Antithetic magnetic and shadow Hamiltonian Monte Carlo," *IEEE Access*, vol. 9, pp. 49857–49867, 2021.

[21] R. Mbuvha, W. T. Mongwe, and T. Marwala, "Separable shadow Hamiltonian hybrid Monte Carlo for Bayesian neural network inference in wind speed forecasting," *Energy AI*, vol. 6, Dec. 2021, Art. no. 100108.

[22] E. Akhmatskaya and S. Reich, "The targeted shadowing hybrid Monte Carlo (TSHMC) method," in *New Algorithms for Macromolecular Simulation*. Berlin, Germany: Springer-Verlag, 2006, pp. 145–158.

[23] T. Radivojević and E. Akhmatskaya, "Modified Hamiltonian Monte Carlo for Bayesian inference," 2017, *arXiv:1706.04032*.

[24] M. Gu and S. Sun, "Neural Langevin dynamical sampling," *IEEE Access*, vol. 8, pp. 31595–31605, 2020.

[25] J. A. Brofos and R. R. Lederman, "Magnetic manifold Hamiltonian Monte Carlo," 2020, *arXiv:2010.07753*.

[26] W. T. Mongwe, R. Mbuvha, and T. Marwala, "Adaptive magnetic Hamiltonian Monte Carlo," *IEEE Access*, vol. 9, pp. 152993–153003, 2021.

[27] E. Hairer, "Backward error analysis for multistep methods," *Numerische Math.*, vol. 84, no. 2, pp. 199–232, Dec. 1999.

[28] W. T. Mongwe, R. Mbuvha, and T. Marwala, "Utilising partial momentum refreshment in separable shadow Hamiltonian hybrid Monte Carlo," *IEEE Access*, vol. 9, pp. 151235–151244, 2021.

[29] W. T. Mongwe, R. Mbuvha, and T. Marwala, "Quantum-inspired magnetic Hamiltonian Monte Carlo," *PLoS ONE*, vol. 16, no. 10, Oct. 2021, Art. no. e0258277.

[30] R. M. Neal, "Probabilistic inference using Markov chain Monte Carlo methods," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, Tech. Rep. CRG-TR-93-1, Jan. 1993.

[31] R. M. Neal, "Bayesian learning via stochastic dynamics," in *Advances in Neural Information Processing Systems*. San Francisco, CA, USA: Morgan Kaufmann, 1993, pp. 475–482. [Online]. Available: http://dl.acm.org/citation.cfm?id=645753.667903

[32] R. Neal, "MCMC using Hamiltonian dynamics," in *Handbook of Markov Chain Monte Carlo*, S. Brooks, A. Gelman, G. L. Jones, and X. L. Meng, Eds. London, U.K.: Chapman & Hall, 2011, pp. 116–162.

[33] J. Sohl-Dickstein, M. Mudigonda, and M. DeWeese, "Hamiltonian Monte Carlo without detailed balance," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 719–726.

[34] M. Betancourt, "A conceptual introduction to Hamiltonian Monte Carlo," 2017, *arXiv:1701.02434*.

[35] H. M. Afshar, R. Oliveira, and S. Cripps, "Non-volume preserving Hamiltonian Monte Carlo and no-u-turnsamplers," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 1675–1683.

[36] E. Hairer, M. Hochbruck, A. Iserles, and C. Lubich, "Geometric numerical integration," *Oberwolfach Rep.*, vol. 3, no. 1, pp. 805–882, 2006.

[37] K.-C. Chen, "Noncanonical Poisson brackets for elastic and micromorphic solids," *Int. J. Solids Struct.*, vol. 44, no. 24, pp. 7715–7730, Dec. 2007.

[38] R. D. Skeel and D. J. Hardy, "Practical construction of modified Hamiltonians," *SIAM J. Scientific Comput.*, vol. 23, no. 4, pp. 1172–1188, Jan. 2001.

[39] A. M. Horowitz, "A generalized guided Monte Carlo algorithm," *Phys. Lett. B*, vol. 268, no. 2, pp. 247–252, Oct. 1991.

[40] D. Vats, J. M. Flegal, and G. L. Jones, "Multivariate output analysis for Markov chain Monte Carlo," *Biometrika*, vol. 106, no. 2, pp. 321–337, Jun. 2019.

[41] L. Kish, "Survey sampling. HN29, K5," Wiley, New York, NY, USA, Tech. Rep. 4, 1965.

[42] A. Gelman and D. B. Rubin, "Inference from iterative simulation using multiple sequences," *Statist. Sci.*, vol. 7, pp. 457–472, Nov. 1992.

[43] V. Roy, "Convergence diagnostics for Markov chain Monte Carlo," *Annu. Rev. Statist. Appl.*, vol. 7, no. 1, pp. 387–412, Mar. 2020.

**WILSON TSAKANE MONGWE** was born in Tembisa, Gauteng, South Africa. He received the B.Sc. degree in computing from the University of South Africa, and the Bachelor of Business Science (B.Bus.Sci.) degree in actuarial science and the master's degree in mathematical finance from the University of Cape Town. He was a recipient of the Google Ph.D. Fellowship in machine learning, which supports his current Ph.D. research with the University of Johannesburg. His research interests include Bayesian machine learning and Markov chain Monte Carlo methods.

**RENDANI MBUVHA** was born in Venda, Limpopo, South Africa. He received the B.Sc. degree (Hons.) in actuarial science and statistics from the University of Cape Town, and the M.Sc. degree in machine learning from KTH, Royal Institute of Technology, Sweden. He is currently a Senior Lecturer in statistics and actuarial science with the University of Witwatersrand, Johannesburg. He is a Qualified Actuary and a Holder of the Chartered Enterprise Risk Actuary Designation. He was a recipient of the Google Ph.D. Fellowship in machine learning, which supported his Ph.D. research with the University of Johannesburg.

**TSHILIDZI MARWALA** (Senior Member, IEEE) was born in Duthuni, Venda, Limpopo, South Africa, in July 1971. He received the bachelor's degree in mechanical engineering from Case Western Reserve University, Cleveland, OH, USA, in 1995, the master's degree in mechanical engineering from the University of Pretoria, Pretoria, South Africa, in 1997, and the Ph.D. degree in engineering from the University of Cambridge (St. Johns College), Cambridge, U.K., in 2000. He is currently a Registered Professional Engineer (Pr. Eng.) with the Engineering Council of South Africa. He has also completed other leadership courses from the Columbia Business School, National University of Singapore, GIBS University of Pretoria, Harvard Business School, and the University of South Africa. Some of his accomplishments include being a fellow of Cambridge Commonwealth Trust, in 1997, CSIR, in 2005, South African Academy of Engineering, in 2007, TWAS, The World Academy of Science, in 2010, African Academy of Science, in 2013, and South African Institute of Electrical Engineers, in 2016. He has won many awards, including notably, the Bronze Order of Mapungubwe awarded by the President of the Republic of South Africa, in 2004, and the TWAS-AAS-Microsoft Award for Young Scientists, in 2009. His contributions in artificial intelligence field come in forms of over 15 books he has authored, over 50 peer-reviewed chapters, over 50 journal publications and over 150 conference publications. He also holds the Deputy Chairperson of the Presidential Fourth Industrial Revolution Commission of South Africa position and has contributed to over 50 articles to local press and journals on the subject of Fourth Industrial Revolution and Economics. He is currently the Vice-Chancellor and the Principal of the University of Johannesburg.