Non-Volume Preserving Hamiltonian Monte Carlo and No-U-Turn Samplers

Hadi Mohasel Afshar[1,2] **Rafael Oliveira**[1,3] **Sally Cripps**[1,2]
[1] ARC Centre for Data Analytics for Resources and Environments, Australia
[2] School of Mathematics and Statistics, the University of Sydney
[3] School of Computer Science, the University of Sydney

Abstract

Volume preservation is usually regarded as a necessary property for the leapfrog transition functions that are used in Hamiltonian Monte Carlo (HMC) and No-U-Turn (NUTS) samplers to guarantee convergence to the target distribution. In this work we rigorously prove that with minimal algorithmic modifications, both HMC and NUTS can be combined with transition functions that are not necessarily volume preserving. In light of these results, we propose a non-volume preserving transition function that conserves the Hamiltonian better than the baseline leapfrog mechanism, on piecewise-continuous distributions. The resulting samplers do not require any assumptions on the geometry of the discontinuity boundaries, and our experimental results show a significant improvement upon traditional HMC and NUTS.

1 Introduction

Hamiltonian Monte Carlo (HMC) (Duane et al., 1987) along with its ingenious variant, the No-U-Turn sampler (NUTS) (Hoffman and Gelman, 2014) are among the most popular Markov chain Monte Carlo (MCMC) samplers (Neal, 2011; Carpenter et al., 2017). These methods utilise gradient information of the target distribution to propose a next state that is reasonably distant from the current state but (under some assumptions) is still accepted with high probability. This is achieved by evolving the current state via a transition function that (a) approximately conserve the Hamiltonian and (b) exactly *preserves volume* i.e. the magnitude of the Jacobian determinant of the transition equals to 1. Poor Hamiltonian approximation results in a lower acceptance probability of the proposal, and therefore reduces the convergence rate, while *volume preservation* has to be exact for convergence to the correct invariant distribution (Neal, 2011).

Levy et al. (2018) propose a variation of HMC that uses a transition function based on neural networks that does not always preserve volume. To account for the lack of volume preservation, Levy et al. (2018) include the magnitude of the Jacobian of the transition in the proposal acceptance probability, as in *Reversible jump Markov chain Monte Carlo* (RJMCMC) (Green, 1995).[1] However Levy et al. (2018) do not formally prove why the inclusion of the Jacobian determinant is required. That is, they do not spell out the correspondence between RJMCMC and HMC.

In this paper, we provide a rigorous proof that directly deals with HMC setting and shows why the inclusion of the Jacobian determinant in the acceptance probability is necessary. Our proof does not require the measure-theoretic aspects of Green's proof and can easily be extended to NUTS, which is a more complex setting.

To illustrate the power of HMC with non-volume preserving transitions, we focus on the family of piecewise continuous distributions (Afshar, 2016). Such distributions emerge in many applications, including: bounded support priors or hard constraints (Lan et al., 2014), non-parametric models based on mixture of truncated basis functions (e.g. piecewise polynomials or exponentials) (Cobb et al., 2006), context-specific conditional densities (such as tree-CPDs) (Sanner and

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

[1] RJMCMC deals with Metorpolis-Hastings proposals for jumping between sub-spaces of different dimensionality via introducing an auxiliary vector to match the dimensionality followed by a bijective transition. It is shown that the magnitude of the Jacobian determinant of such a transition should be included in the proposal acceptance probability to ensure convergence of the sampler to the target.

Abbasnejad, 2012), probabilistic programming (due to *if-then* statements and operations such as *min/max*) (Vajda, 2014), influence diagrams (Howard and Matheson, 2005), piecewise likelihoods as in Preference Learning (Afshar et al., 2015) or Bayesian Belief Update models (Nishimura et al., 2020).

The performance of baseline HMC for exploring piecewise continuous distributions can be quite poor. The reason is that HMC simulates the Hamiltonian dynamics via a leapfrog discretisation that relies on the assumption that the target distribution is smooth. Since piecewise distributions do not satisfy this assumption, the leapfrog simulation can lead to arbitrarily large approximation errors.

Several works have tried to address this shortcoming: Pakman and Paninski (2014) propose methods for HMC-based sampling from piecewise distributions by solving the Hamiltonian equations of motion in their closed form rather than relying on approximations. They conserve the Hamiltonian at the discontinuous boundaries via modifying the momentum similar to the movement of a particle in a physical system. Unfortunately, Hamiltonian equations can rarely be solved exactly. Therefore, the application of this method is restricted to piecewise Gaussian distributions.

Lan et al. (2014) map a constrained/truncated target distribution to a sphere such that the support boundary (in the original space) is mapped to the sphere's equator. This change of domains implicitly handles the original constraints. That is, while an HMC-based sampler moves freely on the sphere, it proposes states that are within the constraints imposed on the original space. The application of this method is restricted to distributions with a single truncated segment.

The least restricted existing variation of HMC that targets piecewise continuous distributions is *Reflective/Refractive HMC* (RHMC) (Afshar and Domke, 2015). RHMC is based on a volume-preserving extension of the leapfrog mechanism that can be applied to arbitrary piecewise continuous distributions with affine discontinuity boundaries. Similar to Pakman and Paninski (2014) method, RHMC needs extra information on the inclination of the boundary hyper-planes.

We propose a simple, easy-to-implement, non-volume preserving transition function, called *Fixed-Orientation Momentum Adjusting Leapfrog* (FORMAL), which approximately conserves the Hamiltonian on piecewise continuous distributions without requiring extra information or assumptions on the geometry of the boundaries. Our experimental results show that the resulting NoVoP HMC and NUTS outperform their volume-preserving counterparts on piecewise distributions.

2 Hamiltonian Monte Carlo (HMC)

Consider a probability density function over $\mathbb{R}^n$: $\pi_{\mathbf{Q}}(\mathbf{q}) \propto \exp(-U(\mathbf{q}))$, where $\mathbf{q}$ and U are referred to as the *position vector* and the *potential energy function* (or simply *energy*), respectively. Starting from the current *position state* $\mathbf{q}$, HMC algorithm (Neal, 2011), proceeds as follows:
(a) A *phase state*, $\mathbf{z} := (\mathbf{q}, \mathbf{p})$, is generated by augmenting the state with an auxiliary *momentum* vector,

$$\mathbf{p} \sim \pi_{\mathbf{P}}(\mathbf{p}) := \mathcal{N}(\mathbf{0}_n, \mathbf{I}_{n\times n}).$$

(b) A deterministic, *volume-preserving* transition function, $\mathcal{F}$ maps $\mathbf{z}$ to a proposed state, $\mathbf{z}' := (\mathbf{q}', \mathbf{p}')$. This transition has a *reversible dynamics* in the sense that,

$$\mathcal{F}(\mathcal{F}(\mathbf{z})) = \mathbf{z}, \qquad \forall \mathbf{z}$$

(c) The proposed state $\mathbf{z}'$ is accepted with probability,

$$\alpha(\mathbf{z}\rightarrow\mathbf{z}') = \min\left\{1, \tfrac{\pi_{\mathbf{Z}}(\mathbf{z}')}{\pi_{\mathbf{Z}}(\mathbf{z})}\right\} = \min\left\{1, \frac{e^{-H(\mathbf{z}')}}{e^{-H(\mathbf{z})}}\right\} \quad (1)$$

where $\pi_{\mathbf{Z}}(\mathbf{z}) = \pi_{\mathbf{P}}(\mathbf{p})\pi_{\mathbf{Q}}(\mathbf{q}) = \exp(-H(\mathbf{z}))$ and the *Hamiltonian*, $H(\cdot)$, is defined as follows:

$$H(\mathbf{z}) := \mathbf{p}^\top \mathbf{p}/2 + U(\mathbf{q}).$$

In baseline HMC, $\mathcal{F}$ simulates the state evolution via the Hamiltonian equations of motion,

$$\frac{d\mathbf{q}}{dt} = \mathbf{p} \qquad \frac{d\mathbf{p}}{dt} = -\nabla U(\mathbf{q}),$$

using L leapfrog discretisation steps, each for a fixed time interval, ϵ. The Hamiltonian is conserved under the evolution defined by the equations of motion. Therefore, if the leapfrog approximation error is low, $H(\mathbf{z}') \approx H(\mathbf{z})$ and the acceptance probability in (1) is high. The Hamiltonian equations of motion as well as their leapfrog approximation preserve volume, in the sense that the total volume of any region in the phase space that is evolved by means of such dynamics, remains constant. Volume preservation is a necessary property for the convergence of baseline HMC sampler to the correct invariant distribution (Neal, 2011). In the next section we present a generalisation that guarantees convergence to the desired distribution regardless of the volume preservation property of the transition function.

3 Non-Volume Preserving HMC (NoVoP HMC)

The proposed *Non-Volume Preserving HMC* (NoVoP HMC) is illustrated in Algorithm 1. If the blue characters are omitted, Algorithm 1 reduces to baseline HMC

Algorithm 1: NovopHMC

```
input  : $\mathbf{q}^{(t)}$, current state (of size $n$); $U$, energy;
         $L$, no. leapfrog steps; $\epsilon$, leapfrog step size
output : next state
$\mathbf{q} \leftarrow \mathbf{q}^{(t)}$;    $\mathbf{p} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{I}_{n\times n})$
$H_0 \leftarrow \|\mathbf{p}\|^2/2 + U(\mathbf{q})$
$J_{\mathcal{F}} \leftarrow 1$          # Jacobian det. of the total transition
for $l = 1$ to $L$ do
    $(\mathbf{q}, \mathbf{p}, |J_\tau|) \leftarrow$ TransitionStep$(\mathbf{q}, \mathbf{p}, \epsilon, U)$
    $J_{\mathcal{F}} \leftarrow J_{\mathcal{F}} \cdot |J_\tau|$
$\mathbf{p} \leftarrow -\mathbf{p}$          # To make the dynamics reversible
$H \leftarrow \|\mathbf{p}\|^2/2 + U(\mathbf{q})$
if $s \sim \mathcal{U}(0,1) < J_{\mathcal{F}} \cdot e^{H_0 - H}$ then
    return q else return $\mathbf{q}^{(t)}$
```

(to facilitate the comparison). In baseline HMC the TransitionStep evolves the phase state for a small time interval, ϵ, via a (volume preserving) leap-frog discretisation step. In NoVoP HMC, TransitionStep is an arbitrary bijective function that does not necessarily preserve volume and instead, along with the evolved state, returns the *magnitude of the Jacobian determinant* of the transition step, $|J_\tau|$. For brevity, throughout we refer to the *magnitude of the Jacobian determinant* as the *Jacobian*.

A concrete example of such TransitionStep will be presented in Section 4 (see Algorithm 2). In the rest of this section we prove that NoVoP HMC converges to the correct target distribution.

Note that the Jacobian of the composition of a series of functions $\tau_1, \ldots, \tau_m$ equals the product of the Jacobians of the individual functions: $|J_{\tau_1 \circ \tau_2 \circ \ldots \circ \tau_m}| = |J_{\tau_1}| \cdot |J_{\tau_2}| \cdots |J_{\tau_m}|$. As such, in Algorithm 1, the value $J_{\mathcal{F}}$ is the Jacobian of the total evolution from a current state, $\mathbf{z} := (\mathbf{q}, \mathbf{p})$, to a proposal state, $\mathbf{z}' := (\mathbf{q}', \mathbf{p}')$, and the acceptance probability is:

$$\alpha(\mathbf{z} \to \mathbf{z}') := \min\left\{1, \left|\frac{\partial \mathbf{z}'}{\partial \mathbf{z}}\right| \cdot \exp\left(H(\mathbf{z}) - H(\mathbf{z}')\right)\right\}$$
$$= \min\left\{1, \left|\frac{\partial \mathbf{z}'}{\partial \mathbf{z}}\right| \cdot \frac{\pi_{\mathbf{Z}}(\mathbf{z}')}{\pi_{\mathbf{Z}}(\mathbf{z})}\right\} \qquad (2)$$

which allows us to derive the following result.

Theorem 3.1. *An ergodic chain of samples drawn via HMC with a transition, $\mathcal{F}$, that has a reversible dynamics, i.e. $\mathcal{F}(\mathcal{F}(\mathbf{z})) = \mathbf{z}$, but is not necessarily volume-preserving, converges to a correct equilibrium distribution if the proposal acceptance probability is decided by* (2).

Proof. Since sampling $\mathbf{p}$ from its marginal distribution leaves the joint distribution invariant, under the assumption of ergodicity of the MCMC chain (Tweedie, 1975), a sufficient condition for the correct convergence is to prove detailed balance for accepting the proposed state, $\mathbf{z}' := \mathcal{F}(\mathbf{z})$, with the probability given by (2), where $\mathbf{z} := (\mathbf{q}, \mathbf{p})$ is the current state. Since the evolution is deterministic, the proposal distribution is in the form of a Dirac delta:

$$Q(\mathbf{z} \to \mathbf{z}') = \delta(\mathbf{z}' - \mathcal{F}(\mathbf{z}))\,.$$

We prove detailed balance condition as given by Green (1995) for the generalised HMC. That is, we show:

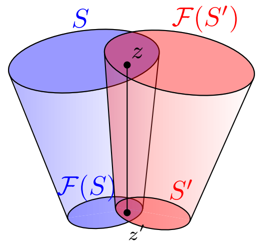
$$\int_{\mathbf{z}\in S} \pi_{\mathbf{Z}}(\mathbf{z}) \int_{\mathbf{z}'\in S'} Q(\mathbf{z} \to \mathbf{z}')\alpha(\mathbf{z} \to \mathbf{z}') d\mathbf{z}' d\mathbf{z}$$
$$= \int_{\mathbf{z}'\in S'} \pi_{\mathbf{Z}}(\mathbf{z}') \int_{\mathbf{z}\in S} Q(\mathbf{z}' \to \mathbf{z})\alpha(\mathbf{z}' \to \mathbf{z}) d\mathbf{z} d\mathbf{z}' \qquad (3)$$

where S and S' are arbitrary volumes (i.e. Borel sets in the parameter space). Intuitively speaking, (3) indicates that to satisfy detailed balance, there should be an equilibrium between the transition from (the probability mass associated with) S to S' and vice versa. As such, we should prove that:

$$\int_{\mathbf{z}\in S} \pi_{\mathbf{Z}}(\mathbf{z}) \int_{\mathbf{z}'\in S'} \delta(\mathbf{z}' - \mathcal{F}(\mathbf{z}))\alpha(\mathbf{z} \to \mathbf{z}') d\mathbf{z}' d\mathbf{z}$$
$$= \int_{\mathbf{z}'\in S'} \pi_{\mathbf{Z}}(\mathbf{z}') \int_{\mathbf{z}\in S} \delta(\mathbf{z} - \mathcal{F}(\mathbf{z}'))\alpha(\mathbf{z}' \to \mathbf{z}) d\mathbf{z} d\mathbf{z}'. \qquad (4)$$

Let $\mathcal{F}(S)$ and $\mathcal{F}(S')$ be the projection of elements of S and S' via $\mathcal{F}$, respectively:

$$\mathcal{F}(S) := \{\mathcal{F}(\mathbf{z}) \mid \mathbf{z} \in S\}$$
$$\mathcal{F}(S') := \{\mathcal{F}(\mathbf{z}') \mid \mathbf{z}' \in S'\}.$$



Since S and S' are arbitrarily chosen, in many cases the two sides of (4) are 0 and the integration bounds can be tightened:
If $\mathbf{z} \in S$ but $\mathcal{F}(\mathbf{z}) \notin S'$ then the LHS of (4) is 0 because for all $\mathbf{z}' \in S'$,

$$\delta(\mathbf{z}' - \mathcal{F}(\mathbf{z})) = 0.$$

On the other hand, if $\mathcal{F}(\mathbf{z}) \in S'$ then by definition $\mathbf{z} \in \mathcal{F}(S')$ since the dynamics of $\mathcal{F}$ is reversible i.e. $\mathbf{z} = \mathcal{F}(\mathcal{F}(\mathbf{z}))$. Therefore, as illustrated in the right hand side diagram, the transition between $\mathbf{z} \in S$ and $\mathbf{z}' \in S'$ is possible iff:

$$\mathbf{z} \in S \cap \mathcal{F}(S') \qquad \text{and} \qquad \mathbf{z}' \in S' \cap \mathcal{F}(S)\,. \qquad (5)$$

As such the integral bounds in (4), can be tightened:

$$\begin{aligned}\int_{\mathbf{z}\in S\cap\mathcal{F}(S')} \pi_{\mathbf{Z}}(\mathbf{z}) \int_{\mathbf{z}'\in S'} \delta(\mathbf{z}'-\mathcal{F}(\mathbf{z}))\alpha(\mathbf{z}\to\mathbf{z}')d\mathbf{z}'d\mathbf{z} \\ = \int_{\mathbf{z}'\in S'\cap\mathcal{F}(S)} \pi_{\mathbf{Z}}(\mathbf{z}') \int_{\mathbf{z}\in S}\delta(\mathbf{z}-\mathcal{F}(\mathbf{z}'))\alpha(\mathbf{z}'\to\mathbf{z})d\mathbf{z}d\mathbf{z}'\,.\end{aligned} \tag{6}$$

Due to the sifting property of the Dirac delta, (6) simplifies to:

$$\int_{\mathbf{z}\in S\cap\mathcal{F}(S')} \pi_{\mathbf{Z}}(\mathbf{z})\alpha(\mathbf{z}\to\mathcal{F}(\mathbf{z}))d\mathbf{z} = \int_{\mathbf{z}'\in S'\cap\mathcal{F}(S)} \pi_{\mathbf{Z}}(\mathbf{z}')\alpha(\mathbf{z}'\to\mathcal{F}(\mathbf{z}'))d\mathbf{z}'\,. \tag{7}$$

By substituting the acceptance probability given by (2) into (7):

$$\begin{aligned}\int_{\mathbf{z}\in S\cap\mathcal{F}(S')} \pi_{\mathbf{Z}}(\mathbf{z}) \min\left\{1, \left|\frac{\partial\mathcal{F}(\mathbf{z})}{\partial\mathbf{z}}\right|\cdot\frac{\pi_{\mathbf{Z}}(\mathcal{F}(\mathbf{z}))}{\pi_{\mathbf{Z}}(\mathbf{z})}\right\} d\mathbf{z} \\ = \int_{\mathbf{z}'\in S'\cap\mathcal{F}(S)} \pi_{\mathbf{Z}}(\mathbf{z}') \min\left\{1, \left|\frac{\partial\mathcal{F}(\mathbf{z}')}{\partial\mathbf{z}'}\right|\cdot\frac{\pi_{\mathbf{Z}}(\mathcal{F}(\mathbf{z}'))}{\pi_{\mathbf{Z}}(\mathbf{z}')}\right\} d\mathbf{z}'\,.\end{aligned} \tag{8}$$

Therefore, to prove that the detailed balance condition holds for non-volume preserving HMC, it is sufficient to show that the two sides of (8) are equal.

Let us define:

$$C := \left\{\mathbf{z} \text{ s.t. } \mathbf{z}\in S\cap\mathcal{F}(S') \text{ and } \left|\frac{\partial\mathcal{F}(\mathbf{z})}{\partial\mathbf{z}}\right|\frac{\pi_{\mathbf{Z}}(\mathcal{F}(\mathbf{z}))}{\pi_{\mathbf{Z}}(\mathbf{z})} < 1\right\}$$

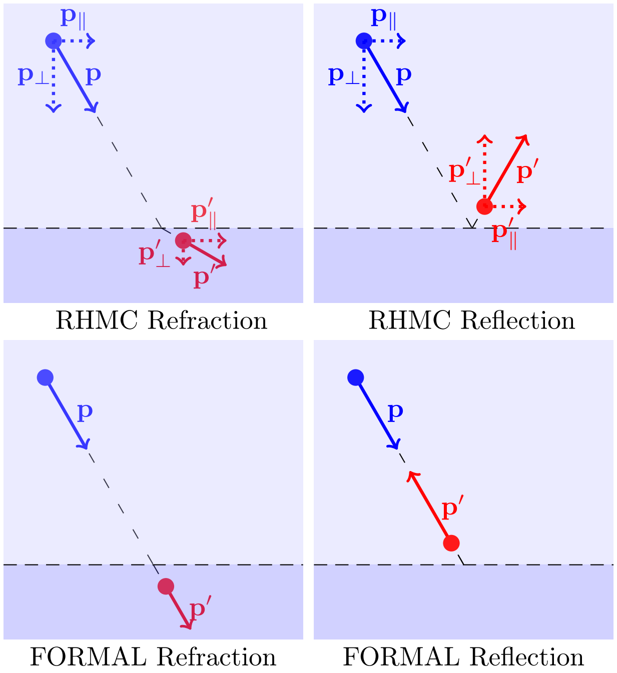
By segmenting $S\cap\mathcal{F}(S')$ to C and its complement $C' := (S\cap\mathcal{F}(S'))\setminus C$, the LHS of (8) becomes:

$$\int_{\mathbf{z}\in C} \pi_{\mathbf{Z}}(\mathbf{z}) \left|\frac{\partial\mathcal{F}(\mathbf{z})}{\partial\mathbf{z}}\right| \cdot \frac{\pi_{\mathbf{Z}}(\mathcal{F}(\mathbf{z}))}{\pi_{\mathbf{Z}}(\mathbf{z})}\, d\mathbf{z} + \int_{\mathbf{z}\in C'} \pi_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z} \tag{9}$$

On the other hand, by a change of random variable $\mathbf{z} := \mathcal{F}(\mathbf{z}')$ and noting that, due to the reversibility of the dynamics, $\mathbf{z}' = \mathcal{F}(\mathbf{z})$ and $\mathbf{z}' \in S'\cap\mathcal{F}(S)$ iff $\mathbf{z}\in S\cap\mathcal{F}(S')$, the RHS of (8) becomes:

$$\begin{aligned}\int_{\mathbf{z}\in S\cap\mathcal{F}(S')} \pi_{\mathbf{Z}}(\mathcal{F}(\mathbf{z})) \min\left\{1, \left|\frac{\partial\mathbf{z}}{\partial\mathcal{F}(\mathbf{z})}\right| \frac{\pi_{\mathbf{Z}}(\mathbf{z})}{\pi_{\mathbf{Z}}(\mathcal{F}(\mathbf{z}))}\right\} \left|\frac{\partial\mathcal{F}(\mathbf{z})}{\partial\mathbf{z}}\right| d\mathbf{z} \\ = \int_{\mathbf{z}\in C} \pi_{\mathbf{Z}}(\mathcal{F}(\mathbf{z})) \left|\frac{\partial\mathcal{F}(\mathbf{z})}{\partial\mathbf{z}}\right| d\mathbf{z} \\ + \int_{\mathbf{z}\in C'} \pi_{\mathbf{Z}}(\mathcal{F}(\mathbf{z})) \left|\frac{\partial\mathbf{z}}{\partial\mathcal{F}(\mathbf{z})}\right| \cdot \frac{\pi_{\mathbf{Z}}(\mathbf{z})}{\pi_{\mathbf{Z}}(\mathcal{F}(\mathbf{z}))} \cdot \left|\frac{\partial\mathcal{F}(\mathbf{z})}{\partial\mathbf{z}}\right| d\mathbf{z}\end{aligned} \tag{10}$$

Now, it can easily be verified that (9) and (10) are equal, which completes the proof. □



RHMC Refraction RHMC Reflection

FORMAL Refraction FORMAL Reflection

Figure 1: Evolution of a position state (blue circle) with momentum $\mathbf{p} := \mathbf{p}_{\perp} + \mathbf{p}_{\parallel}$ into a position vector (red circle) with momentum $\mathbf{p}' := \mathbf{p}'_{\perp} + \mathbf{p}'_{\parallel}$ in the adjacency of a discontinuity boundary (horizontal dashed line) of a piecewise-continuous target via RHMC refraction/reflection (first row) and FORMAL refraction/reflection (second row) transitions.

4 Fixed-Orientation Momentum Adjusting Leapfrog (FORMAL)

So far we have proved that the sampling scheme in Algorithm 1 will converge to draws from the invariant distribution. However, to maximise the rate of convergence, the proposal acceptance probability given by (2) should be close to 1. One way to achieve this is to require that the transition should approximately preserve both volume and Hamiltonian, i.e. $|\partial\mathcal{F}(\mathbf{z})/\partial\mathbf{z}| \approx 1$ and $H(\mathcal{F}(\mathbf{z})) \approx H(\mathbf{z})$.

In this section, we propose such a transition, namely, *Fixed-Orientation Momentum Adjusting Leapfrog* (FORMAL), which outperforms the leapfrog mechanism on piecewise models. The latter algorithm is blind to the boundaries of piecewise models, and through passing a discontinuity boundary, it encounters a Hamiltonian approximation error equal to the energy difference on the two sides of the boundary, ΔU.

In an existing extension of the leapfrog mechanism, used in Reflective-HMC (RHMC) (Afshar and Domke, 2015), to restore the Hamiltonian when a discontinuity boundary is met, the state's momentum is decomposed to components $\mathbf{p}_\perp$ and $\mathbf{p}_\parallel$ that are perpendicular and parallel to the boundary, respectively. If $\|\mathbf{p}_\perp\|^2 > 2\Delta U$, then the state *refracts* by substituting the perpendicular component $\mathbf{p}_\perp$ with $\mathbf{p}'_\perp := \sqrt{\|\mathbf{p}_\perp\|^2 - 2\Delta U} \cdot \frac{\mathbf{p}_\perp}{\|\mathbf{p}_\perp\|}$, otherwise it *reflects* by changing its direction $\mathbf{p}'_\perp := -\mathbf{p}_\perp$ (Figure 1, first row).

To guarantee volume preservation, RHMC can only be applied to piecewise distributions with affine boundaries. Furthermore, in practice, the required momentum decomposition may be hard or impossible to compute since it needs the normal vector i.e. a vector that is orthogonal to the boundary surface at any point where a discontinuity boundary is met.

In FORMAL, this problematic decomposition step is omitted: If $\|\mathbf{p}\|^2 > 2\Delta U$, the Hamiltonian is restored by modifying the momentum's magnitude. $\mathbf{p}' := \sqrt{\|\mathbf{p}\|^2 - 2\Delta U} \cdot \frac{\mathbf{p}}{\|\mathbf{p}\|}$. Otherwise we reverse its direction, $\mathbf{p}' := -\mathbf{p}$ (Figure 1, second row). In general, this transition does not preserve volume. However, it does preserve the Hamiltonian and leads to a reversible dynamics.[2]

Algorithm 2 illustrates a transition step that in the time-interval ϵ, can reflect or refract multiple times based on the FORMAL mechanism.[3] This algorithm only needs to compute the Jacobian, $|J_\tau|$, in case refractions occur because the other steps of the algorithm preserve volume. In our implementation, we have used the JAX automatic differentiation framework (Bradbury et al., 2018) to automate the computation of the Jacobians.

Algorithm 2: TRANSITIONSTEP

```
input  : $(\mathbf{q}, \mathbf{p})$, state; $U(\cdot)$, energy; $\epsilon$, step size.
output : evolved state and its Jacobian
         determinant w.r.t. the input state.
$|J_\tau| = 1$   # Jacobian det. of the total transition step
$\mathbf{p} \leftarrow \mathbf{p} - \epsilon \nabla U(\mathbf{q})/2$   # Half-step
# Full-step:
$\langle \mathbf{x}, t_x, \Delta U \rangle \leftarrow$ FIRSTHIT$(\mathbf{q}, \mathbf{p}, \epsilon, U)$
if $\langle \mathbf{x}, t_x, \Delta U \rangle = \emptyset$ then
    $\mathbf{q} \leftarrow \mathbf{q} + \epsilon \mathbf{p}$
else
    # Keep trace of $\mathbf{q}$ and $\mathbf{p}$ before their non-volume
      preserving evolution:
    $(\mathbf{q}^*, \mathbf{p}^*) \leftarrow (\mathbf{q}, \mathbf{p})$
    $t_0 \leftarrow 0$
    while $\langle \mathbf{x}, t_x, \Delta U \rangle \neq \emptyset$ do
        $\mathbf{q} \leftarrow \mathbf{x}$
        $t_0 \leftarrow t_0 + t_x$
        if $\|\mathbf{p}\|^2 > 2\Delta U$ then   #Adjust magnitude:
            $\mathbf{p} \leftarrow \sqrt{\|\mathbf{p}\|^2 - 2\Delta U} . \frac{\mathbf{p}}{\|\mathbf{p}\|}$
        else   #Reverse direction:
            $\mathbf{p} \leftarrow -\mathbf{p}$
        $\langle \mathbf{x}, t_x, \Delta U \rangle \leftarrow$ FIRSTHIT$(\mathbf{q}, \mathbf{p}, \epsilon - t_0, U)$
    $\mathbf{q} \leftarrow \mathbf{q} + (\epsilon - t_0)\mathbf{p}$
    $|J_\tau| \leftarrow \left| \det \frac{\partial(\mathbf{q}, \mathbf{p})}{\partial(\mathbf{q}^*, \mathbf{p}^*)} \right|$
$\mathbf{p} \leftarrow \mathbf{p} - \epsilon \nabla U(\mathbf{q})/2$   # Half-step
return $(\mathbf{q}, \mathbf{p}, |J_\tau|)$
Note: FIRSTHIT $(\mathbf{q}, \mathbf{p}, t_{\max}, U)$ is any function
  that returns a tuple $\langle \mathbf{x}, t_x, \Delta U \rangle$, where $\mathbf{x}$ is the
  first intersection of a boundary (of the piecewise
  function $U$) with line segment $[\mathbf{q}, \mathbf{q} + t_{\max}\mathbf{p}]$;
  $t_x := (\mathbf{x} - \mathbf{q})/\mathbf{p}$ is the time it is visited, and $\Delta U$
  is the change of energy at the discontinuity. If no
  such intersection exists, $\emptyset$ is returned.
```

5 No-U-Turn sampler (NUTS)

NUTS sampler (Hoffman and Gelman, 2014) is based on the idea of setting the number of leapfrog steps, L, adaptively by continuing the leapfrog mechanism till the path made by leapfrogs makes a U-turn. That is, the last proposed position state, $\mathbf{q}'$, starts becoming closer to the initial position $\mathbf{q}$. This happens if $(\mathbf{q}' - \mathbf{q}) \cdot \mathbf{p}' \leq 0$. Unfortunately, such an approach violates the detailed balance condition. To overcome this problem, in NUTS, a path of states (*traced set* $\mathcal{B}$) is generated such that its size is somehow decided by the mentioned termination criterion. The set $\mathcal{B}$ has the following important property, which is required for detailed balance:

$$P(\mathcal{B}|\mathbf{z}) = P(\mathcal{B}|\mathbf{z}') \qquad \forall \mathbf{z}, \mathbf{z}' \in \mathcal{B}\,. \tag{11}$$

That is, the probability of generating $\mathcal{B}$, starting from any of its members is the same. Having $\mathcal{B}$, the current state $\mathbf{z}$ and a *slice variable*, u, that is uniformly drawn from the interval $[0, \pi_{\mathbf{Z}}(\mathbf{z})]$, a set of *chosen states* $\mathcal{C}$:

$$\mathcal{C} := \{\mathbf{z}' \in \mathcal{B} \text{ s.t. } \pi_{\mathbf{Z}}(\mathbf{z}') \geq u\}, \tag{12}$$

is constructed from which the next state is drawn uniformly.

5.1 Limitations of NUTS

If the leapfrog transition is not effective in preserving the Hamiltonian (which in particular, is the case with piecewise continuous distributions), the described algo-

[2] Algorithm 2 does not have a reversible dynamics by itself. However, if it maps a state $(\mathbf{q}, \mathbf{p})$ to $(\mathbf{q}', \mathbf{p}')$, then it also maps $(\mathbf{q}', -\mathbf{p}')$ to $(\mathbf{q}, -\mathbf{p})$. As such, when used as a sub-routine within Algorithm 1, the dynamics of the total transition becomes reversible since in line 7 of Algorithm 1, $\mathbf{p} \rightarrow -\mathbf{p}$.

[3] If the blue characters are omitted, the algorithm reduces to a baseline leapfrog step.

rithm can generate a huge or even infinite[4] $\mathcal{B}$ containing several states that will not be chosen (i.e. $||\mathcal{B}|| >> ||\mathcal{C}||$) and waste computational resources. To prevent this issue, the designers of NUTS have proposed an extra criterion. Terminate if:

$$u \geq \exp(\Delta_{\max} - H(\mathbf{z}')) \tag{13}$$

where $\Delta_{\max}$ is a positive parameter set by the user. It can be verified that the probability of termination via criterion (13) is non-zero if $H(\mathbf{z}') > H(\mathbf{z}) + \Delta_{\max}$. That is, the approximation error in the simulation of Hamiltonian dynamics is more than $\Delta_{\max}$. However, the following proposition entails that this criterion violates detailed balance.

Proposition 1. *If $\Delta_{\max} < \infty$, termination based on criterion (13) violates the intended condition (11).*[5]

To minimise the interference with detailed balance, the designers of NUTS propose to choose a large value, e.g. $\Delta_{\max} := 1000$ (Hoffman and Gelman, 2014). However, we do not find this solution systematic, since (a) if $\Delta_{\max}$ is set to be too large, the interference with detailed balance is minimized at the expense of low performance (due to wasteful expansion of $\mathcal{B}$), while (b) if $\Delta_{\max}$ is set smaller, the performance improves by sacrificing the correct convergence. And (c) the magnitude of $\Delta_{\max}$ only makes sense relative to the normalisation constant of the target distribution, which is often unknown.

In the next section we show that FORMAL dynamics can neatly be combined with NUTS. The result is a non-volume preserving version of NUTS, which (a) does not rely on the extra parameter $\Delta_{\max}$, (b) perfectly satisfies detailed balance, and (c) has a better performance, since it preserves the Hamiltonian better.

6 Non-Volume Preserving No-U-Turn sampler (NoVoP NUTS)

Algorithms 3 and 4 describe NoVoP NUTS and, if the blue parts are omitted, they reduce to baseline NUTS. In both versions of NUTS, to return a sample from a (position) distribution, $\exp(-U(\mathbf{q}))$, given a current sample, $\mathbf{q}$, a phase state, $\mathbf{z} := (\mathbf{q}, \mathbf{p})$, is made by drawing a same-sized momentum vector from a standard normal distribution. Therefore, similar to HMC, we have $\pi_{\mathbf{Z}}(\mathbf{z}) := \exp(-\frac{1}{2}\mathbf{p}^{\top}\mathbf{p} - U(\mathbf{q}))$. A *slice variable*, u, is uniformly drawn from $[0, \pi_{\mathbf{Z}}(\mathbf{z})]$. A *trace path* $\mathcal{B}$ and a set of *chosen states* $\mathcal{C} \subset \mathcal{B}$ are generated by starting from $\mathcal{B} = \{\mathbf{z}\}$ and iteratively doubling its size by adding new states to its right (or left) side via forward (resp. backward) evolution of the rightmost (resp. leftmost) member of the path, till a termination criterion is satisfied. This process is such that the property given by (11) is satisfied (as long as $\Delta_{\max} = \infty$). The trace path, $\mathcal{B}$, is not explicitly generated, but can be thought of as the initial state plus the set of all states that are traced by BldTree, and a subset, $\mathcal{C}'$, of them is added to the *chosen states*, $\mathcal{C}$.

Unlike the baseline, in NoVoP NUTS the Jacobian of each chosen state, $\mathbf{z}' \in \mathcal{C}$, w.r.t. the current state, $\mathbf{z}$, (i.e. $\left|\frac{\partial \mathbf{z}'}{\partial \mathbf{z}}\right|$) is computed. These Jacobians are built recursively by multiplying the Jacobian of the last transition step in the already computed Jacobian associated with the right-most (or left-most) traced state. As such, per transition step, only the Jacobian of that step is computed, and redundant computations are avoided. Furthermore, in NoVoP NUTS, a state, $\mathbf{z}' \in \mathcal{B}$, is added to $\mathcal{C}$ if $u \leq \left|\frac{\partial \mathbf{z}'}{\partial \mathbf{z}}\right| \pi_{\mathbf{Z}}(\mathbf{z}')$, while in baseline NUTS, the condition is: $u \leq \pi_{\mathbf{Z}}(\mathbf{z}')$ (see Line 4 of Alg. 4). In NoVoP NUTS, the probability of adding a state, $\mathbf{z}'$, to $\mathcal{C}$, given the trace path, $\mathcal{B}$, and slice variable, u, is then:

$$P(\mathbf{z}' \in \mathcal{C} | \mathcal{B}, u) := \begin{cases} \mathbb{I}\left[u \leq \pi_{\mathbf{Z}}(\mathbf{z}') \cdot \left|\frac{\partial \mathbf{z}'}{\partial \mathbf{z}}\right|\right] & \text{if } \mathbf{z}' \in \mathcal{B} \\ 0 & \text{otherwise} \end{cases}$$

while in the baseline, the Jacobian term is missing. As the last difference between the two algorithms, note that NoVoP HMC does not rely on the termination condition given by (13) since in this algorithm, we set $\Delta_{\max} = \infty$ which is equivalent to omitting (the red) Line 8 in Alg. 4.

Theorem 6.1. *An ergodic chain of samples drawn via NoVoP NUTS sampling with an arbitrary (not necessarily volume-preserving) bijective transition step converges to the correct invariant distribution.*[6]

7 Experimental results

As discussed, NoVoP HMC (with FORMAL transition) is designed for piecewise continuous distributions where the existing volume preserving extension of HMC (namely, RHMC) (Afshar and Domke, 2015) is impossible or difficult to apply. Nonetheless, it is constructive to first consider a distribution where both NoVoP HMC and RHMC can be applied. Clearly we would expect the volume preserving technique to perform better in this situation. However, it is informative to measure the difference between these approaches in practice.

Section 7.1 compares NoVoP HMC against RHMC on a piecewise model with affine boundaries.

[4] For instance, on a piecewise constant (e.g. uniform) distribution, the momentum is invariant under the leapfrog mechanism. Therefore, a U-turn never occurs.

[5] The proof and a counterexample to correct convergence are provided in the supplementary material.

[6] The proof provided in the supplementary material.

Algorithm 3: NoVoPNUTS

```
Input: $\mathbf{q}^{(t)}$, current state (of size $n$);
leapfrog step size $\epsilon$
$\mathbf{p} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{I}_{n\times n})$
$u \sim \mathrm{Unif}[0, \exp\left(-\frac{1}{2}\mathbf{p}^\top\mathbf{p} - U(\mathbf{q}^{(t)})\right)]$
$(\mathbf{q}^-, \mathbf{p}^-) \leftarrow (\mathbf{q}^{(t)}, \mathbf{p})$
$(\mathbf{q}^+, \mathbf{p}^+) \leftarrow (\mathbf{q}^{(t)}, \mathbf{p})$
$J^- \leftarrow 1$;    $J^+ \leftarrow 1$            # Jacobian determinants
$j \leftarrow 0$;   $\mathcal{C} \leftarrow \left\{(\mathbf{q}^{(t)}, \mathbf{p}, 1)\right\}$;   $s \leftarrow 1$
while $s = 1$ do
    $v_j \sim \mathrm{Unif}(\{-1, 1\})$            # Choose a direction
    if $v_j = -1$ then
        $\mathbf{q}^-, \mathbf{p}^-, J^-, -, -, -, \mathcal{C}', s' \leftarrow$
          BldTree$(\mathbf{q}^-, \mathbf{p}^-, J^-, u, v_j, j, \epsilon)$
    else
        $-, -, -, \mathbf{q}^+, \mathbf{p}^+, J^+, \mathcal{C}', s' \leftarrow$
          BldTree$(\mathbf{q}^+, \mathbf{p}^+, J^+, u, v_j, j, \epsilon)$
    if $s' = 1$ then
        $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{C}'$
    if $(\mathbf{q}^+ - \mathbf{q}^-) \cdot \mathbf{p}^- < 0$ or $s' = 0$ or
          $(\mathbf{q}^+ - \mathbf{q}^-) \cdot \mathbf{p}^+ < 0$ then  $s \leftarrow 0$
    $j \leftarrow j + 1$
return $(\mathbf{q}^{(t+1)}, -, -)$ uniformly at random from $\mathcal{C}$
```

Algorithm 4: BldTree

```
Input: $(\mathbf{q}, \mathbf{p})$, state; $J$, Jacobian; $u$, slice variable;
$v$, direction; $j$, tree depth; $\epsilon$, step size; $\Delta_{\max} := \infty$
if $j = 0$ then
    $(\mathbf{q}', \mathbf{p}', |J_\tau|) \leftarrow$ TransitionStep$(\mathbf{q}, \mathbf{p}, v\epsilon)$
    $J' \leftarrow J \cdot |J_\tau|$
    if $u \leq J' \cdot \exp(-\frac{1}{2}\mathbf{p}'^\top\mathbf{p}' - U(\mathbf{q}'))$ then
        $\mathcal{C}' \leftarrow \{(\mathbf{q}', \mathbf{p}', J')\}$
    else
        $\mathcal{C}' \leftarrow \{\}$
    $s' \leftarrow \mathbb{I}[u < \exp\{\Delta_{\max} - U(\mathbf{q}') - \frac{1}{2}\mathbf{p}'^\top\mathbf{p}'\}]$
    return $\mathbf{q}', \mathbf{p}', J', \mathbf{q}', \mathbf{p}', J', \mathcal{C}', s'$
else
    $\mathbf{q}^-, \mathbf{p}^-, J^-, \mathbf{q}^+, \mathbf{p}^+, J^+, \mathcal{C}', s' \leftarrow$
        BldTree $(\mathbf{q}, \mathbf{p}, J, u, v, j-1, \epsilon)$
    if $v = -1$ then
        $\mathbf{q}^-, \mathbf{p}^-, J^-, -, -, -, \mathcal{C}'', s'' \leftarrow$
          BldTree$(\mathbf{q}^-, \mathbf{p}^-, J^-, u, v, j-1, \epsilon)$
    else
        $-, -, -, \mathbf{q}^+, \mathbf{p}^+, J^+, \mathcal{C}'', s'' \leftarrow$
          BldTree$(\mathbf{q}^+, \mathbf{p}^+, J^+, u, v, j-1, \epsilon)$
    $s' \leftarrow s' \cdot s'' \cdot \mathbb{I}[(\mathbf{q}^+ - \mathbf{q}^-) \cdot \mathbf{p}^- \geq 0]$
        $\cdot\, \mathbb{I}[(\mathbf{q}^+ - \mathbf{q}^-) \cdot \mathbf{p}^+ \geq 0]$
    $\mathcal{C}' \leftarrow \mathcal{C}' \cup \mathcal{C}''$
    return $\mathbf{q}^-, \mathbf{p}^-, J^-, \mathbf{q}^+, \mathbf{p}^+, J^+, \mathcal{C}', s'$
```

Section 7.2 presents results for a distribution where the boundaries are not affine, and therefore RHMC cannot be guaranteed to produce draws from the correct invariant distribution. Here, we compare the performance of NoVoP HMC/NUTS against their volume preserving counterparts. All algorithms are implemented in Python and run on a single thread of a 2.7GHz CPU. NoVoP algorithms use JAX auto-differentiation framework (Bradbury et al., 2018) to compute Jacobians.[7]

7.1 Experiment 1

We compare NoVoP HMC against Baseline HMC (Neal, 2011), RHMC, (Afshar and Domke, 2015), and Random-walk Metropolis-Hastings (RWMH) on the model introduced in Afshar and Domke (2015):[8]

$$U(\mathbf{q}) = \begin{cases} \sqrt{\mathbf{q}^\intercal A\, \mathbf{q}}, & \text{if } \|\mathbf{q}\|_\infty \leq 3 \\ 1 + \sqrt{\mathbf{q}^\intercal A\, \mathbf{q}}, & \text{if } 3 \leq \|\mathbf{q}\|_\infty \leq 6 \\ +\infty, & \text{otherwise} \end{cases} \qquad (14)$$

where A is a positive definite matrix that is randomly generated for each simulation. All variations of HMC use same fixed parameters $L = 10$ and $\epsilon = 0.1$.

Due to the symmetry of the model, the ground truth expected value of $\mathbf{q} = (q_1, \ldots, q_n)$ is known to be $\mathbf{0}$. As such, the absolute error of the expected value of q_d is $|\sum_{k=1}^{K} q_d^{(k)}/K|$, for $d = 1, \ldots, n$, where $q_d^{(k)}$ denotes the k^{th} iterate of q_d in an MCMC chain. The worst mean absolute error (WMAE) over all dimensions is taken as the error measurement of the chain:

$$\mathrm{WMAE}\left(\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(K)}\right) = \frac{1}{K} \max_{d=1,\ldots,n} \left| \sum_{k=1}^{K} q_d^{(k)} \right| \qquad (15)$$

For each algorithm, 10 MCMC chains are run, and in each chain the value of each element of the initial state, $\mathbf{q}^{(0)}$, is uniformly drawn from the interval $[5.5, 5.99)$. The average WMAE of all chains ± 95% confidence interval are depicted in Figure 2. In low dimensional settings, $n = 5$, our results show that all samplers have similar convergence rates. However, in high dimensional settings, $n = 20$, the convergence rates of Baseline HMC and RWMH are much slower than those of RHMC and NoVoP HMC. As predicted, the convergence rate of RHMC is better than that of NoVoP HMC, although the difference is marginal.

7.2 Experiment 2

In this experiment, we use the following piecewise continuous target distribution where the boundaries are

[7] Our source code is available at: `https://github.com/hadimafshar/NoVopHMC.git`

[8] A 2-dimensional instance of such a distribution is plotted in the supplementary material.

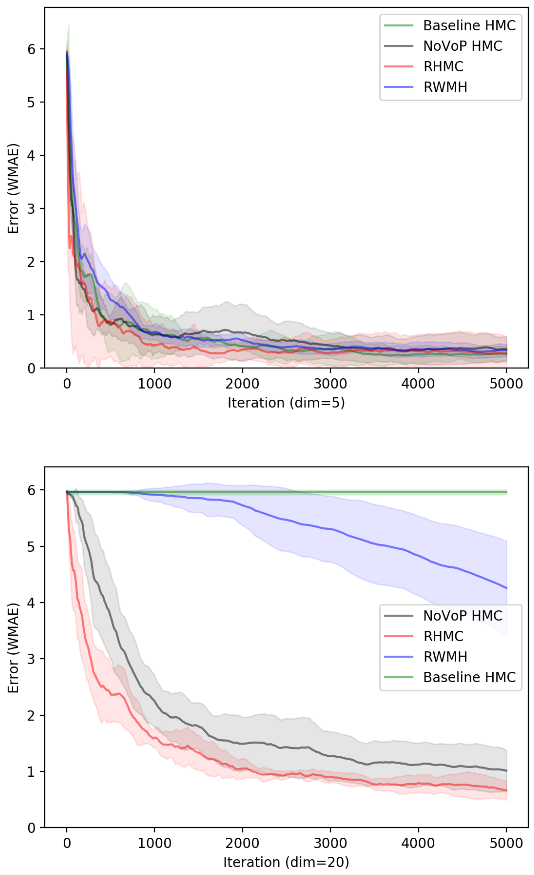


Figure 2: The worst mean absolute error (WMAE), see (15), vs no. samples drawn from the model given by (14) with dimensionality 5 (top) and 20 (bottom).

hyper-spherical (in contrast to the hyper-cubic boundaries of the previous experimental model):[9]

$$U(\mathbf{q}) = \begin{cases} \sqrt{\mathbf{q}^\intercal A\, \mathbf{q}}, & \text{if } \|\mathbf{q}\|_2 \leq 3 \\ 1 + \sqrt{\mathbf{q}^\intercal A\, \mathbf{q}}, & \text{if } 3 \leq \|\mathbf{q}\|_2 \leq 6 \\ 50 + \sqrt{\mathbf{q}^\intercal A\, \mathbf{q}}, & \text{otherwise} \end{cases} \tag{16}$$

Since the boundaries are non-affine, we omit RHMC from this experiment but add the baseline NUTS and NoVoP NUTS samplers. For baseline NUTS we fix $\epsilon = 0.1$ and $\Delta_{\max} = 1000$, while for NoVoP NUTS, we fix $\epsilon = 0.1$ and $\Delta_{\max} = \infty$ (as discussed in Section 6). In both variations of NUTS, we restrict the maximum size of the traced set $\mathcal{B}$ to 2^{12}.[10]

Figure 3 depicts WMAE versus the number of drawn samples (top) and the sampling time (bottom) where the model dimensionality is set to $n = 50$. For each sampling algorithm, except baseline NUTS, 10 MCMC chains are run for 5000 iterations. In each MCMC chain, each element of the initial sample vector is uniformly drawn from the interval $[5.5/\sqrt{n}, 5.9/\sqrt{n})$. For baseline NUTS we only run 1 chain and stop it after 2000 iterations, since this algorithm is at least two orders of magnitude slower than other samplers. While all other samplers return 5000 samples in less than 20 minutes, baseline NUTS could not draw more than 2000 samples after 24 hours.[11]

These plots show that in terms of convergence versus iterations, as well as convergence versus time, NoVoP HMC performs the best. Since this method relies on extra computations of Jacobians, such a high performance versus time may seem counter-intuitive. However, note that the only non-volume preserving operations are *refractions* that make a small subset of total transitional operations and do not affect the speed much.[12]

Figure 4 provides a better insight on the performance of the compared samplers. In this figure, for each sampling algorithm, a sample trance of a single MCMC chain is plotted versus the sampling iterations and time. According to these plots, on the presented piecewise model, baseline HMC has a high rejection rate. This issue is fixed in NoVoP HMC via a better preservation of the Hamiltonian which leads to a significant increase the acceptance probability of the proposals. Baseline and NoVoP NUTS algorithms, do not have an explicit proposal acceptance step and their performance can be better evaluated versus time. It can be seen that baseline NUTS sampling is extremely slow (which is due to the formation of very large trace sets $\mathcal{B}$). Although NoVoP NUTS is much faster than the baseline NUTS, it still halts sporadically for long periods of time to form very large trace sets, and in terms of error versus time, at least on this model, does not perform as good as NoVoP HMC.

[9] Some 2-dimensional instances of such a distribution are plotted in the supplementary material.

[10] We have intentionally chosen a high maximum size to highlight the difference between HMC and NUTS. By reducing this limit, the performance of NUTS/NoVoP-NUTS become more similar to HMC/NoVoP-HMC.

[11] For baseline NUTS, we also tried setting ϵ adaptively via the *dual averaging* mechanism proposed in Hoffman and Gelman (2014). However, due to the problem of this sampler with piecewise models, the adaptively chosen ϵ became too close to 0, which slowed down the sampler furthermore and deteriorated its performance.

[12] In our experimental models, by increasing the dimensionality, the rate of refraction drops while the rate of reflection usually increases. For example in Experiment 2 (with $\epsilon = 0.1$ and $L = 10$), for dimensionalities $n = 5$, 10 and 50, the average number of refractions per sample are respectively 0.14, 0.12 and 0.09, while the average number of reflections per sample are respectively 0.10, 0.22 and 1.9.

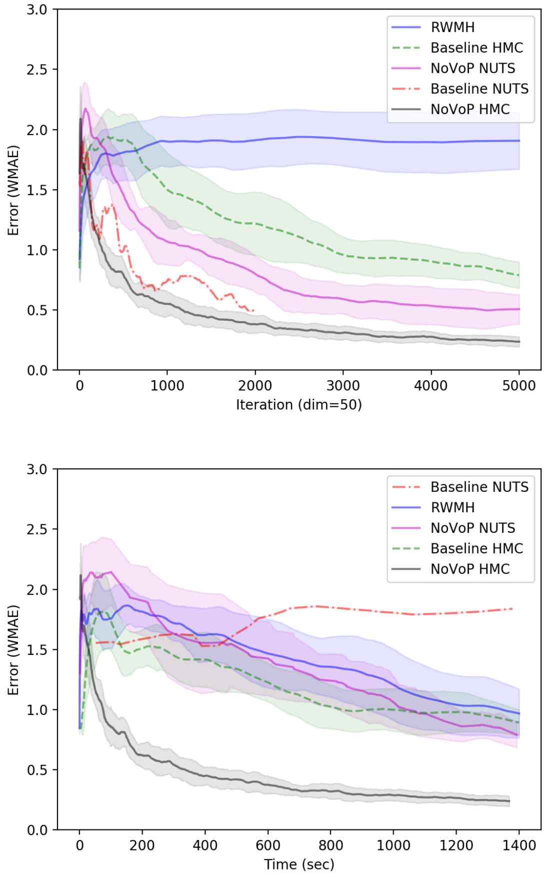


Figure 3: WMAE vs the number of samples (top) and sampling time in seconds (bottom).

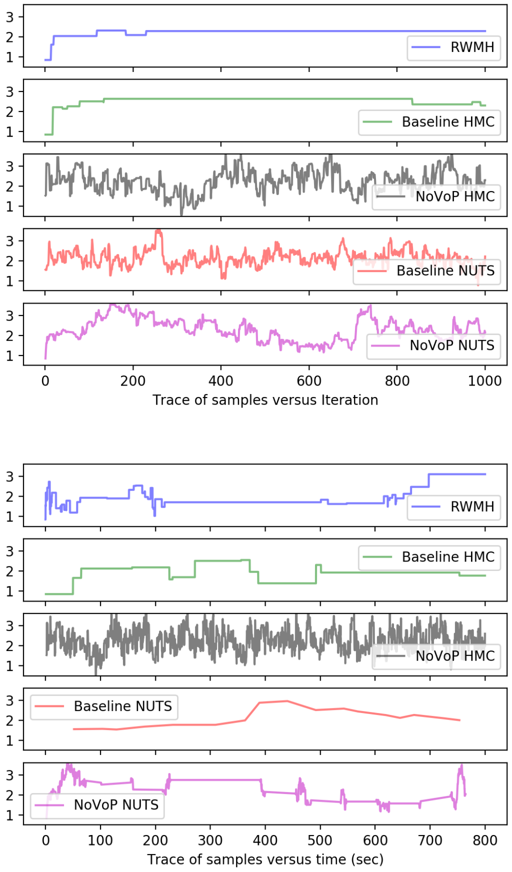


Figure 4: Trace of samples drawn from a single MCMC chain versus iterations (top) and time (bottom).

8 Conclusion

In this paper, we propose a generalisation of HMC and NUTS sampling techniques to arbitrary transition functions that have reversible dynamics but are not necessarily volume preserving. As mentioned, an instance of such a sampler is already proposed in the literature (Levy et al., 2018) but a rigorous proof for its correct convergence has been missing. We provided proofs that directly target HMC and NUTS settings and indicate that with the inclusion of the Jacobian in the proposal acceptance probability, non-volume preserving HMC and NUTS converge to the correct equilibrium distribution. This result can potentially lead to a variety of new samplers with transitions that do not have the limitations of the existing leapfrog mechanism and seek a trade-off between preservation of Hamiltonian and volume, rather than sacrificing the former to guarantee the exact preservation of the latter.

We also propose a specific transition mechanism, namely, *Fixed-Orientation Momentum Adjusting Leapfrog* (FORMAL), which is suitable for piecewise continuous distributions and is not limited to affine boundaries. Furthermore it does not require extra information on the inclination of the boundaries, and its implementation is easy, thanks to existing auto-differentiation frameworks, which automate the computation of the required Jacobians. As our experimental results show, the extra computational overhead does not have a significant effect on the sampling speed, while the better preservation of the Hamiltonian can lead to a significant improvement in the sampling scheme when compared to the baseline.

References

Afshar, H. M. (2016). *Probabilistic inference in piecewise graphical models*. PhD thesis, The Australian National University (Australia).

Afshar, H. M. and Domke, J. (2015). Reflection, refraction, and Hamiltonian Monte Carlo. In *Advances in neural information processing systems*, pages 3007–3015.

Afshar, H. M., Sanner, S., and Abbasnejad, E. (2015). Linear-time Gibbs sampling in piecewise graphical models. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., and Wanderman-Milne, S. (2018). JAX: composable transformations of Python+NumPy programs.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).

Cobb, B. R., Shenoy, P. P., and Rumí, R. (2006). Approximating probability density functions in hybrid bayesian networks with mixtures of truncated exponentials. *Statistics and Computing*, 16(3):293–308.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics letters B*, 195(2):216–222.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Howard, R. A. and Matheson, J. E. (2005). Influence diagrams. *Decision Analysis*, 2(3):127–143.

Lan, S., Zhou, B., and Shahbaba, B. (2014). Spherical hamiltonian monte carlo for constrained target distributions. In *International Conference on Machine Learning*, pages 629–637. PMLR.

Levy, D., Hoffman, M. D., and Sohl-Dickstein, J. (2018). Generalizing hamiltonian monte carlo with neural networks. *International Conference on Learning Representations (ICLR)*.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., editors, *Handbook of Markov Chain Monte Carlo*, pages 113–162. Chapman & Hall.

Nishimura, A., Dunson, D. B., and Lu, J. (2020). Discontinuous Hamiltonian Monte Carlo for discrete parameters and discontinuous likelihoods. *Biometrika*, 107(2):365–380.

Pakman, A. and Paninski, L. (2014). Exact Hamiltonian Monte Carlo for truncated multivariate gaussians. *Journal of Computational and Graphical Statistics*, 23(2):518–542.

Sanner, S. and Abbasnejad, E. (2012). Symbolic variable elimination for discrete and continuous graphical models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26.

Tweedie, R. L. (1975). Sufficient conditions for ergodicity and recurrence of markov chains on a general state space. *Stochastic Processes and their Applications*, 3(4):385–403.

Vajda, S. (2014). *Probabilistic programming*. Academic Press.