Orbital MCMC

Kirill Neklyudov
University of Amsterdam
(currently at Vector Institute)
k.necludov@gmail.com

Max Welling
University of Amsterdam
CIFAR

Abstract

Markov Chain Monte Carlo (MCMC) algorithms ubiquitously employ complex deterministic transformations to generate proposal points that are then filtered by the Metropolis-Hastings-Green (MHG) test. However, the condition of the target measure invariance puts restrictions on the design of these transformations. In this paper, we first derive the acceptance test for the stochastic Markov kernel considering arbitrary deterministic maps as proposal generators. When applied to the transformations with orbits of period two (involutions), the test reduces to the MHG test. Based on the derived test we propose two practical algorithms: one operates by constructing periodic orbits from any diffeomorphism, another on contractions of the state space (such as optimization trajectories). Finally, we perform an empirical study demonstrating the practical advantages of both kernels.

1 INTRODUCTION

MCMC is an ubiquitous computational tool across many different scientific fields such as statistics, bioinformatics, physics, chemistry, machine learning, etc. Generation of the proposal samples for continuous state spaces usually includes sophisticated deterministic transitions. The most popular example of such transitions is the evolution of the Hamiltonian dynamics (Duane et al., 1987; Hoffman & Gelman, 2014). More recently, the combination of neural models and conventional MCMC algorithms has attracted the community's attention (Song et al., 2017; Hoffman et al., 2019), and such hybrid models already found their application in the problems of modern physics (Kanwar et al., 2020; Albergo et al., 2021). These developments motivate further studies of the possible benefits of deterministic functions in MCMC methods.

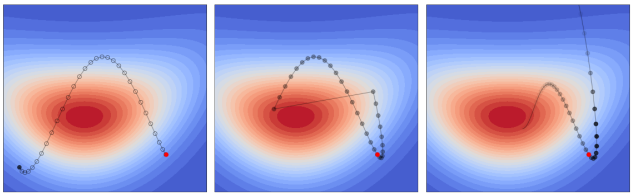

Figure 1: The illustration of the proposed algorithms. The transparency of a point corresponds to its weight. Red dots correspond to the initial states. From left to right: HMC (stochastically accepts only the last state), Orbital kernel on a periodic orbit (accepts all the states weighted), Orbital kernel on an infinite orbit (accepts the states within a certain region).

One possible way to introduce a deterministic transition into the MCMC kernel is to consider an involutive function inside a stochastic transition kernel as described in (Neklyudov et al., 2020; Spanbauer et al., 2020) (we recap Involutive MCMC in Section 2.2). One of the practical benefits of this approach is that it allows one to learn the kernel as a neural model improving its mixing properties (Spanbauer et al., 2020).

Our paper extends prior works by deriving novel transition kernels that do not rely on involutions. We start by deriving the novel acceptance test that is applicable for any diffeomorphism and operates on its orbits. However, in its general form, the test might be infeasible to compute, requiring the evaluation of the density across the orbit, which might have an infinite number of points. To obtain practical algorithms we consider two special cases. Our first algorithm operates on periodic orbits, and we demonstrate how one can easily design such orbits by introducing auxiliary variables (Section 3). The second algorithm operates on infinite orbits of optimization trajectories. We demonstrate

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

how such orbits could be truncated, preserving the target measure (see section 4). Finally, we provide an empirical study of these algorithms and discuss their advantages and disadvantages (Section 5).

2 BACKGROUND

2.1 Mean ergodic theorem

The core idea of the MCMC framework is the mean ergodic theorem, namely, its adaptation to the space of distributions. That is, consider the ϕ-irreducible Markov chain (see Roberts et al. (2004)) with the kernel $k(x' \,|\, x)$ that keeps the distribution p invariant. Denoting the single step of the chain as an operator $K : [Kq](x') = \int dx \; k(x' \,|\, x) q(x)$, the mean ergodic theorem says that its time average projects onto the invariant subspace:

$$\lim_{n\to\infty} \left\| \frac{1}{n} \sum_{i=0}^{n-1} K^i p_0 - p \right\| = 0, \quad \text{for all} \;\; p_0. \tag{1}$$

Note, that the irreducibility condition implies the uniqueness of the invariant distribution. In practice, the application of K is not tractable, though. Therefore, one resorts to Monte Carlo estimates by accepting a set of samples $\{x_0, \dots, x_{n-1}\}$, each coming with a weight $1/n$, and distributed as $x_i \sim [K^i p_0](x)$. From this perspective, the MCMC designs and analyses the kernels that allow for projections, as in (1).

2.2 Involutive MCMC

In this section, we briefly discuss the main idea of the Involutive MCMC (iMCMC) framework (Neklyudov et al., 2020; Spanbauer et al., 2020), which is the initial point of our further developments. The iMCMC framework starts by considering the kernel

$$\begin{aligned} k(x' \,|\, x) =& \delta(x' - f(x)) g(x) + \delta(x' - x)(1 - g(x)) \\ g(x) =& \min\left\{1, \frac{p(f(x))}{p(x)} \left| \frac{\partial f}{\partial x} \right| \right\}, \end{aligned} \tag{2}$$

where $p(x)$ is the target density, $\delta(x)$ is the Dirac delta-function and $f(x)$ is a diffeomorphism. The stochastic interpretation of this kernel is the following. Starting at the point x the chain accepts $f(x)$ with the probability $g(x)$ or stays at the same point with the probability $(1 - g(x))$.

For the kernel (2) to preserve the target measure $[Kp](x') = \int dx \; k(x' \,|\, x) p(x) = p(x')$, we have

$$\begin{aligned} &\min\left\{ p(f^{-1}(x)) \left| \frac{\partial f^{-1}}{\partial x} \right|, p(x) \right\} = \\ &= \min\left\{ p(x), p(f(x)) \left| \frac{\partial f}{\partial x} \right| \right\} \quad \forall x. \end{aligned} \tag{3}$$

Algorithm 1 Involutive MCMC

input target density $p(x)$
input density $p(v \,|\, x)$ and a sampler from $p(v \,|\, x)$
input involutive $f(x, v) : f(x, v) = f^{-1}(x, v)$
  initialize x
  for $i = 0 \dots n$ **do**
    sample $v \sim p(v \,|\, x)$
    propose $(x', v') = f(x, v)$
    $g(x, v) = \min\{1, \frac{p(x', v')}{p(x, v)} \left| \frac{\partial f(x, v)}{\partial [x, v]} \right| \}$
    $x_i = \begin{cases} x', & \text{with probability } g(x, v) \\ x, & \text{with probability } (1 - g(x, v)) \end{cases}$
    $x \leftarrow x_i$
  end for
output samples $\{x_0, \dots, x_n\}$

This equation allows one to bypass the difficulties of designing a measure-preserving function by choosing f such that $f^{-1}(x) = f(x)$ or, equivalently $f(f(x)) = x$. Thus, the iMCMC framework proposes a practical way to design f that implies the invariance of $p(x)$ by considering the family of involutions. However, inserting such f into the transition kernel (2) reduces it to jump only between two points: from x to $f(x)$ and then back to $f(f(x)) = f^{-1}(f(x)) = x$.

To be able to cover the support of the target distribution with involutive f, the iMCMC framework introduces an additional source of stochasticity into (2) through auxiliary variables. That is, instead of traversing the target $p(x)$, the chain traverses the distribution $p(x, v) = p(x)p(v \,|\, x)$, where $p(v \,|\, x)$ is an auxiliary distribution that we are free to choose. The key ingredients for choosing $p(v \,|\, x)$ are easy computation of its density and the ability to efficiently sample from it. Then, interleaving the kernel (2) with the resampling of the auxiliary variable $v \,|\, x$, one can potentially reach any state $[x, v]$ of the target distribution $p(x, v)$ (see Algorithm 1). Samples from the marginal distribution of interest $p(x)$ can now simply be obtained by ignoring the dimensions that correspond to the variable v.

It turns out that by choosing different involutions f and auxiliary distributions $p(v \,|\, x)$ in Algorithm 1 one can formulate a large class of MCMC algorithms as described in (Neklyudov et al., 2020). However, the involutive property of the considered deterministic map f enforces the iMCMC kernel (Algorithm 1) to be reversible:

$$k(x', v' \,|\, x, v) p(x, v) = k(x, v \,|\, x', v') p(x', v'). \tag{4}$$

Moreover, the marginalized kernel on x: $\widehat{k}(x' \,|\, x) = \int dv dv' \;\; k(x', v' \,|\, x, v) p(v \,|\, x)$ is also reversible: $\widehat{k}(x' \,|\, x) p(x) = \widehat{k}(x \,|\, x') p(x')$, which might hinder the mixing properties of the chain. To design

irreversible kernels, the iMCMC framework proposes to compose several reversible kernels.

3 ORBITAL KERNEL

3.1 Derivation of the acceptance test

Starting with the acceptance test (2) the iMCMC framework looks for a family of suitable deterministic maps f, and ends up with involutive functions (see Section 2.2). In this work, we approach the problem of kernel design differently: given a deterministic map f we develop a kernel that admits the target density as its eigenfunction, generalizing the iMCMC framework.

We start our reasoning with the kernel (which we call the escaping orbital kernel) $k(x' \,|\, x) = \delta(x' - f(x))g(x) + \delta(x' - x)(1 - g(x))$, where $g(x)$ is the acceptance test, and the kernel tries to make a move along the trajectory of $f(x)$ as in iMCMC. Putting this kernel into the condition $Kp = p$: $\int dx\; k(x' \,|\, x)p(x) = p(x')$, we have

$$\begin{aligned}\int dy\; \delta(x' - y)g(f^{-1}(y))p(f^{-1}(y))\left|\frac{\partial f^{-1}}{\partial y}\right| + \\ + p(x') - \int dx\; \delta(x' - x)g(x)p(x) = p(x'),\end{aligned} \tag{5}$$

where in the first term we have applied a change of variables to $y = f(x)$. This simplifies to

$$g(x')p(x') = g(f^{-1}(x'))p(f^{-1}(x'))\left|\frac{\partial f^{-1}}{\partial x}\right|_{x=x'}. \tag{6}$$

Let's assume that x' is an element of the orbit $\mathrm{orb}(x_0)$, which we can describe using the notation of iterated functions ($f^0(x) = \mathrm{id}(x) = x, f^{n+1}(x) = f(f^n(x))$) as $\mathrm{orb}(x_0) = \{f^k(x_0), k \in \mathbb{Z}\}$. Then putting $x' = f^k(x_0),\ k \in \mathbb{Z}$ into (6) we get

$$\begin{aligned}\forall k \in \mathbb{Z} \quad g(f^k(x_0))p(f^k(x_0))\left|\frac{\partial f^k}{\partial x}\right|_{x=x_0} = \\ = g(f^{k-1}(x_0))p(f^{k-1}(x_0))\left|\frac{\partial f^{k-1}}{\partial x}\right|_{x=x_0}.\end{aligned} \tag{7}$$

Thus, instead of a single equation (6) at point x' we actually have a system of equations generated by the recursive application of (7): $\forall k \in \mathbb{Z}$

$$g(x_0)p(x_0) = g(f^k(x_0))p(f^k(x_0))\left|\frac{\partial f^k}{\partial x}\right|_{x=x_0}. \tag{8}$$

The system tells us that the function f must preserve the measure $g(x)p(x)$ on the orbit $\mathrm{orb}(x_0)$. Note that if f preserves the target $p(x)$ then any constant $g(x) = c$ satisfies this equation. In this case, we can set $g(x) = 1$ and use iterated applications of f for sampling from the target distribution (ensuring that the chain could densely cover the state space). However, the design of non-trivial measure-preserving f (especially with dense orbits) is infeasible in most cases. Therefore, we can assume that $f(x)$ preserves some other density $q(x)$ on the orbit $\mathrm{orb}(x_0)$:

$$q(x_0) = q(f^k(x_0))\left|\frac{\partial f^k}{\partial x}\right|_{x=x_0} \quad \forall k \in \mathbb{Z}. \tag{9}$$

From (8), we also know that f must preserve $g(x)p(x)$. Thus, we can think of $g(x)$ as an importance weight that makes $g(x)p(x) \propto q(x)$. Interpreting $g(x)$ as a probability ($0 \le g(x) \le 1$), we have

$$0 \le g(f^k(x_0)) = c \cdot \frac{q(f^k(x_0))}{p(f^k(x_0))} \le 1, \;\; \forall k \in \mathbb{Z}. \tag{10}$$

where the constant c makes sure that $g(x) \le 1$. This is precisely achieved by $c = \inf_{k\in\mathbb{Z}} \left\{\frac{p(f^k(x))}{q(f^k(x))}\right\}$ for any $x \in \mathrm{orb}(x_0)$. Finally, using the formula for c and equations (9), (10) at $x = f^k(x_0)$, the test function $g(x)$ for any $x \in \mathrm{orb}(x_0)$ is

$$g(x) = \frac{q(x)}{p(x)} \inf_{k\in\mathbb{Z}} \left\{\frac{p(f^k(x))}{q(f^k(x))}\right\} = \inf_{k\in\mathbb{Z}} \left\{\frac{p(f^k(x))}{p(x)}\left|\frac{\partial f^k}{\partial x}\right|\right\}.$$

Note that c is the greatest lower bound on $p(x)/q(x)$ for $x \in \mathrm{orb}(x_0)$, but any lower bound guarantees $g(x) \le 1$. The choice of the greatest lower bound is motivated by the stochastic interpretation of the chain, i.e., we want to minimize the probability of staying in the same state. Although there is a rigorous result (Peskun, 1973) for the comparison of chains via the probabilities of staying in the same state, it applies only for reversible kernels, which is not always applicable to the kernel under consideration.

Putting the test back into the kernel we have the following result.

Theorem 1. *(Escaping orbital kernel)*
Consider a target density $p(x)$ and continuous bijective function f. The transition kernel

$$\begin{aligned}k(x' \,|\, x) =& \delta(x' - f(x))g(x) + \delta(x' - x)(1 - g(x)) \\ g(x) =& \inf_{k\in\mathbb{Z}} \left\{\frac{p(f^k(x))}{p(x)}\left|\frac{\partial f^k}{\partial x}\right|\right\}\end{aligned} \tag{11}$$

keeps the target density invariant

$$[Kp](x') = \int dx\; k(x' \,|\, x)p(x) = p(x').$$

If we would like to accept several points from the same orbit we can substitute f with f^m for any integer m and get the following test.

Corollary 1. *For the the kernel* $k_m(x' \,|\, x) = \delta(x' - f^m(x))g_m(x) + \delta(x' - x)(1 - g_m(x))$ *that satisfies* $K_m p = p$*, the maximal acceptance probability* $g_m(x)$ *is*

$$g_m(x) = \inf_{k \in \mathbb{Z}} \left\{ \frac{p(f^{mk}(x))}{p(x)} \left| \frac{\partial f^{mk}}{\partial x} \right| \right\}. \tag{12}$$

Once we have designed the acceptance test we can check if it is reversible. Reversibility may hinder the mixing properties of the kernel and is usually considered as an undesirable property (Turitsyn et al., 2011). In the following proposition we see that the derived test theoretically allows us to design irreversible kernels.

Proposition 1. *(Reversibility criterion)*
Consider the escaping orbital kernel $k(x' \,|\, x) = \delta(x' - f(x))g(x) + \delta(x' - x)(1 - g(x))$ *that satisfies* $Kp = p$*, and* $g(x) > 0$*. This kernel is reversible w.r.t.* p*, i.e.* $k(x' \,|\, x)p(x) = k(x \,|\, x')p(x')$*, if and only if* f *is an involution, i.e.,* $f(x) = f^{-1}(x)$*.*

Proof. See Appendix A.3.1. □

The bottleneck of the proposed practical scheme is the evaluation of the test. Indeed, it requires the evaluation of the infimum over the whole orbit, and if we don't know its analytical formula its computation becomes infeasible. However, as we demonstrate further, some types of orbits may simplify this evaluation.

The orbit $\mathrm{orb}(x_0)$ is *periodic* if there is an integer $T > 0$ such that for any $x \in \mathrm{orb}(x_0)$, and for any $k \in \mathbb{Z}$, we have $f^{k+T}(x) = f^k(x)$. The *period* of the orbit $\mathrm{orb}(x_0)$ is the minimal $T > 0$ satisfying $f^{k+T}(x) = f^k(x)$. Thus, the orbit $\mathrm{orb}(x_0)$ with the period T can be represented as a finite set of points: $\mathrm{orb}(x_0) = \{f^0(x_0), f^1(x), \ldots, f^{T-1}(x)\}$, and the acceptance test for periodic orbits can be reduced to the minimum over this finite set as follows.

Proposition 2. *(Periodic orbit)*
For periodic orbit $\mathrm{orb}(x)$ *with period* T*, the acceptance test in the kernel* (11) *(from Theorem 1) becomes*

$$\begin{aligned} g(x) &= \inf_{k \in \mathbb{Z}} \left\{ \frac{p(f^k(x))}{p(x)} \left| \frac{\partial f^k}{\partial x} \right| \right\} = \\ &= \min_{k=0,\ldots,T-1} \left\{ \frac{p(f^k(x))}{p(x)} \left| \frac{\partial f^k}{\partial x} \right| \right\}. \end{aligned} \tag{13}$$

For periodic orbits with $T = 2$ (for instance, orbits of involutions), this test immediately yields the Metropolis-Hastings-Green test (or the iMCMC kernel (2)):

$$\min_{k=0,1} \left\{ \frac{p(f^k(x))}{p(x)} \left| \frac{\partial f^k}{\partial x} \right| \right\} = \min \left\{ 1, \frac{p(f(x))}{p(x)} \left| \frac{\partial f}{\partial x} \right| \right\}.$$

By considering other lower bounds in the inequalities (10) for periodic orbits with $T = 2$, one can derive other conventional tests, for instance, the tests induced by Barker's lemma (Barker, 1965). Another special case appears when the points of an orbit come arbitrarily close to the initial point, which we consider in Appendix A.2.

Finally, we need to be able to jump between orbits to be able to cover the state space densely. We can achieve it in the same way as in Algorithm 1, i.e. we can introduce auxiliary variables that allow to jump to another orbit by simply resampling them.

3.2 Accepting the whole orbit

Even for periodic orbits the acceptance test (13) might be inefficient because we need to evaluate the minimum across the whole orbit (and the density is not preserved). However, the evaluation of the target density over the whole orbit allows for accepting several samples from the orbit at once. The naive way to accept several points is to consider a linear combination of the escaping orbital kernels starting from the same point. We discuss it in details in Appendix B.1.

Another way to make use of the evaluated densities over the whole orbit is as follows. Instead of switching to another orbit after the accept/reject step, we may let the kernel continue on the same orbit, and thus collect more samples. Once again, we consider the kernel $k(x' \,|\, x) = \delta(x' - f(x))g(x) + \delta(x' - x)(1 - g(x))$ that preserves target measure $Kp = p$. Starting from the delta-function at some initial point $p_0(x) = \delta(x - x_0)$, the chain iterates by applying the corresponding operator:

$$\begin{aligned} [Kp_0](x') &= \int dx\; k(x' \,|\, x)p_0(x) = \\ &= (1 - g(x_0))\delta(x' - x_0) + g(x_0)\delta(x' - f(x_0)). \end{aligned} \tag{14}$$

Denoting the recurrence relation as $p_{t+1} = Kp_t$, we obtain the sum of weighted delta-functions along the orbit $\mathrm{orb}(x_0)$ of f:

$$p_t(x) = [K^t p_0](x) = \sum_{i=0}^{t} \omega_i^t \delta(x - f^i(x_0)), \tag{15}$$

where ω_i^t is the weight of the delta-function at $f^i(x_0)$ after t steps. Thus, instead of considering the operator K on the space of functions we consider it on the space of sequences, and analyse the limit of the series $1/n \sum_{i=0}^{n-1} K^i$, as in the mean ergodic theorem. The result of our analysis is as follows.

Theorem 2. *(Convergence on a single orbit)*
Consider the proper escaping orbital kernel ($Kp = p$*, and* $g(x) > 0$*) applied iteratively to* $p_0(x) = \delta(x - x_0)$*.*

For aperiodic orbits, iterations yield

$$[K^t p_0](x) = \sum_{i=0}^{t} \omega_i^t \delta(x - f^i(x_0)),$$
$$\lim_{t\to\infty} \sum_{t'=0}^{t} \omega_i^{t'} = \frac{1}{g(f^i(x_0))}, \quad \lim_{t\to\infty} \frac{1}{t} \sum_{t'=0}^{t} \omega_i^{t'} = 0. \tag{16}$$

For periodic orbits with period T, the time average is

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=0}^{t-1} K^i p_0 = \sum_{i=0}^{T-1} \omega_i \delta(x - f^i(x_0)),$$
$$\omega_i = \frac{p(f^i(x_0)) \left|\frac{\partial f^i}{\partial x}\right|_{x=x_0}}{\sum_{j=0}^{T-1} p(f^j(x_0)) \left|\frac{\partial f^j}{\partial x}\right|_{x=x_0}}. \tag{17}$$

Proof. See Appendix A.4. □

Thus, for periodic orbits, the time average of every weight converges to a very reasonable value, which is proportional to the target density at this point. Now, using the weights from (17), we can accept the whole orbit by properly weighting each sample, and we can verify that by accepting the whole orbit we preserve the target measure by considering the kernel $k(x' \,|\, x) = \sum_{i=0}^{T-1} \omega_i \delta(x' - f^i(x))$, where ω_i is given by (17). The practical value of this procedure comes from the fact that the cost of weight evaluations is equivalent to the cost of the test in Theorem 1.

The kernel behaves differently for aperiodic orbits. Indeed, starting from the point x_0, the kernel always moves the probability mass further along the orbit escaping any given point on the orbit since there is no mass flowing backward (the weight of $f^{-1}(x_0)$ is zero from the beginning). Intuitively, we can think that the weights ω_i^t from (15) evolve in time as a wave packet moving on the real line. With this intuition it becomes evident that the time average at every single point on the orbit converges to zero since the mass always escapes this point. Therefore, if we want to accept several points from the orbit, we need to wait until the kernel "leaves" this set of points and then stop the procedure. We derived a formula for this situation, which turns out to be a deterministic version of SNIS (Andrieu et al., 2003), and provide a simple example of this procedure in Appendix B.2.

3.3 Practical algorithm

Based on the previous derivations, we propose a practical sampling scheme. Starting from some initial state x_0 we would like to traverse the orbit $\text{orb}(x_0)$ and accept all the states $\{f^i(x_0)\}_{i\in\mathbb{Z}}$ with the weights ω_i derived in Proposition 2. To guarantee the unbiasedness of the procedure, we design a deterministic function with periodic orbits based on f.

Algorithm 2 Orbital MCMC (periodic)

input target density $p(x)$
input density $p(v \,|\, x)$ and a sampler from $p(v \,|\, x)$
input continuous $f(x, v)$
  initialize $[x, d]$
  for N iterations **do**
    sample $v \sim p(v \,|\, x)$
    collect orbit $\{[x_j, v_j, d_j] = \widehat{f}^j(x, v, d)\}_{j=0}^{T-1}$
    $\omega_i \leftarrow p(f^i(x, v)) \left|\partial f^i / \partial [x, v]\right|$
    $\omega_i \leftarrow \omega_i / \sum_j \omega_j$
    $[x, d] \leftarrow [x_j, d_j]$ with probability ω_j
    samples $\leftarrow$ samples $\cup \{(\omega_i, x_i)\}_{i=0}^{T-1}$
  end for
output samples

Given a function f, one can easily construct a periodic function by introducing the auxiliary discrete variable $d \in \{0, \dots, T-1\}$ (direction), which decides how we apply f to the current state x. That is, we design $\widehat{f}(x, d)$ on the extended space of tuples $[x, d]$ with orbits of period T as follows.

$$\widehat{f}(x, d) = \begin{cases} [f(x), (d+1) \bmod T], \text{ if } d < T-1 \\ [f^{-(T-1)}(x), (d+1) \bmod T], \text{ if } d = T-1 \end{cases}$$

Indeed, starting from any pair $[x, d]$ and iteratively applying $\widehat{f}(x, d)$, we have

$$\widehat{f}^T(x, d) = \widehat{f}^{d+1}\Big([f^{T-1-d}(x), T-1]\Big) =$$
$$= \widehat{f}^d\Big([f^{-d}(x), 0]\Big) = [f^d(f^{-d}(x)), d] = [x, d]. \tag{18}$$

To switch between orbits of $f(x)$ we still need another auxiliary variable to be able to cover the state space densely. As in Algorithm 1, we introduce the auxiliary variable v assuming that we can easily sample $v \sim p(v \,|\, x, d)$ and evaluate the joint density $p(x, v, d)$. Thus, on the switching step, we sample a single point $[x_j, d_j]$ from the orbit interpreting the weights ω_j as probabilities, and then resample the auxiliary variable v. For simplicity, we consider $p(d) = \text{Uniform}\{0, \dots, T-1\}$, and $p(x, v, d) = p(x, v)p(d)$. Gathering all of the steps we get Algorithm 2.

Note that the directional variable forces the chain to perform both "forward" and "backward" moves, which can be considered as a limitation of this algorithm since it may increase the autocorrelation. To alleviate this effect we make the chain irreversible by shifting the direction by $T/2$ after each iteration. In practice, this update performs better than uniform sampling of d.

4 DIFFUSING ORBITAL KERNEL

We start this section by illustrating the limits of the escaping orbital kernel through a simple example. Consider the map $f(x) = x+1$ in $\mathbb{R}$, which has only infinite orbits. Then the test of the escaping kernel

$$g(x) = \inf_{k\in\mathbb{Z}} \left\{ \frac{p(x+k)}{p(x)} \right\} = 0 \tag{19}$$

for any target density $p(x)$. Hence, we can't use the map $f(x) = x + 1$ to sample even a single point from the target distribution because the density goes to zero faster than the Jacobian. At the same time, if we just try to use the formula (17) saying that the period of the orbit is infinite, we end up with a valid kernel. That is, $k(x' \,|\, x) = \sum_{i=-\infty}^{+\infty} \omega_i \delta(x' - f^i(x))$, where $\omega_i = p(f^i(x))/\sum_{j=-\infty}^{+\infty} p(f^j(x))$. Inserting this kernel into $Kp = p$, one can make sure that this kernel indeed keeps the target measure invariant. This example motivates another kernel, which is able to traverse infinite trajectories:

$$\begin{aligned} k(x' \,|\, x) = \delta(x - x')(1 - g^+(x) - g^-(x)) + \\ +\delta(x' - f(x))g^+(x) + \delta(x' - f^{-1}(x))g^-(x). \end{aligned} \tag{20}$$

Writing the condition $Kp = p$ for this kernel, we derive two possible solutions (see Appendix A.6). One of the solutions is a linear combination of escaping orbital kernels that we have already mentioned before (Appendix B.1). Another solution yields a new acceptance test:

$$\begin{aligned} g^+(x) &= p(f(x)) \left| \frac{\partial f}{\partial x} \right| c(x), \\ g^-(x) &= p(f^{-1}(x)) \left| \frac{\partial f^{-1}}{\partial x} \right| c(x), \end{aligned} \tag{21}$$

where $c(x)$ may be chosen as

$$c = \frac{1}{2} \inf_{k\in\mathbb{Z}} \left\{ \frac{1}{p(f^k(x)) \left| \frac{\partial f^k}{\partial x} \right|} \right\}, \quad \text{or}$$

$$c = \inf_{k\in\mathbb{Z}} \left\{ \frac{1}{p(f^{k+1}(x)) \left| \frac{\partial f^{k+1}}{\partial x} \right| + p(f^{k-1}(x)) \left| \frac{\partial f^{k-1}}{\partial x} \right|} \right\}$$

the second one is optimal in the sense of the minimum rejection probability, but we further proceed with the first one for simplicity. Note that taking f as an involution, the kernel (20) becomes equivalent to the iMCMC kernel (2). In Appendix A.7, we prove that the diffusing orbital kernel is always reversible.

Proposition 3. *(Reversibility)*
The diffusing orbital kernel (20) *with test* (21) *is reversible w.r.t. p:* $k(x' \,|\, x)p(x) = k(x \,|\, x')p(x')$.

To provide the reader with the intuition we again assume that the map f preserves some density q on the orbit $\mathrm{orb}(x_0)$, i.e. $q(x_0) = q(f^i(x_0)) \left| \partial f^i / \partial x \right|$. We can then rewrite the test slightly differently:

$$\begin{aligned} g^+(x) = \widehat{g}(f(x)),\ g^-(x) = \widehat{g}(f^{-1}(x)), \text{where} \\ \widehat{g}(x) = \widehat{c} \frac{p(x)}{q(x)},\ \ \widehat{c} = \frac{1}{2} \inf_k \left\{ \frac{q(f^k(x_0))}{p(f^k(x_0))} \right\}. \end{aligned} \tag{22}$$

Firstly, it is now apparent that the diffusing orbital kernel is tightly related to rejection sampling. Indeed, $\widehat{g}(x)$ accepts samples with a probability proportional to the target density, and the constant c makes this probability smaller than 1: $\widehat{c} \le q(x)/(2p(x))$.

Secondly, we can compare this test with the test for the escaping orbital kernel:

$$g(x) = c \cdot \frac{q(x)}{p(x)},\ \ c = \inf_{k\in\mathbb{Z}} \left\{ \frac{p(f^k(x))}{q(f^k(x))} \right\}. \tag{23}$$

We see that both tests are complementary to each other. That is, if $p(x)/q(x)$ vanishes then $c = 0$ and we cannot use the escaping orbital kernel; however, at the same time, $\widehat{c}$ does not go to zero and we can use the diffusing orbital kernel. The same logic applies in the opposite direction when $q(x)/p(x)$ vanishes. In practice, having a continuous bijection f one can decide between these kernels by estimating $\inf\{p(x)/q(x)\}$ and $\inf\{q(x)/p(x)\}$ on the orbit. Finally, the inversion of the density ratio under the infimum allows the diffusing kernel to converge on aperiodic orbits as stated in the following proposition.

Theorem 3. *(Convergence on a single orbit)*
Consider the diffusing orbital kernel (with test (21)*, and* $c > 0$*), and the initial distribution* $p_0(x) = \delta(x - x_0)$*. For aperiodic orbits, if the series* $\sum_{j=-\infty}^{+\infty} p(f^j(x_0)) \left| \frac{\partial f^j}{\partial x} \right|$ *converges, we have*

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=0}^{t-1} K^i p_0 = \sum_{i=-\infty}^{+\infty} \omega_i \delta(x - f^i(x_0)), \tag{24}$$

$$\omega_i = \frac{p(f^i(x_0)) \left| \frac{\partial f^i}{\partial x} \right|_{x=x_0}}{\sum_{j=-\infty}^{+\infty} p(f^j(x_0)) \left| \frac{\partial f^j}{\partial x} \right|_{x=x_0}}. \tag{25}$$

For periodic orbits with period T*, we have*

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=0}^{t-1} K^i p_0 = \sum_{i=0}^{T-1} \omega_i \delta(x - f^i(x_0)), \tag{26}$$

$$\omega_i = \frac{p(f^i(x_0)) \left| \frac{\partial f^i}{\partial x} \right|_{x=x_0}}{\sum_{j=0}^{T-1} p(f^j(x_0)) \left| \frac{\partial f^j}{\partial x} \right|_{x=x_0}}. \tag{27}$$

Proof. See Appendix A.8. □

Algorithm 3 Orbital MCMC (contracting)

input target density $p(x)$
input density $p(v \mid x)$ and a sampler from $p(v \mid x)$
input continuous $f(x, v)$
input threshold value W
  for N iterations **do**
    sample $v \sim p(v \mid x)$
    $\omega_{\max} \leftarrow p(x, v)$
    while $\log \omega_i > \log \omega_{\max} - \log W$ **do**
      $[x_i, v_i] = f^i(x, v)$
      $\omega_i \leftarrow p(f^i(x, v))\left|\partial f^i / \partial [x, v]\right|$
      $\omega_{\max} \leftarrow \max\{\omega_i, \omega_{\max}\}$
    end while
    $\omega_i \leftarrow \omega_i / \sum_j \omega_j$
    $[x] \leftarrow [x_i]$ with probability ω_i
    samples $\leftarrow$ samples $\cup \{(\omega_i, x_i)\}$
  end for
output samples

4.1 Practical algorithm

Although, the derived kernel allows us to sample using the infinite trajectories, it still puts some restrictions on the choice of the diffeomorphism f. The obvious restriction is that the sum $\sum_{j=-\infty}^{+\infty} p(f^j(x_0))\left|\partial f^j / \partial x\right|$ has to converge. One way to make this series convergent is to choose f as the step of an optimization algorithm maximizing the log-probability of the target distribution. For instance, f can be chosen as the gradient ascent

$$f(x) = x + \varepsilon \nabla \log p(x). \tag{28}$$

For $i \to +\infty$, the point converges to some local maximum $f^i(x_0) \to x^*$, hence, for proper target distributions, the density converges to some positive constant $p(f^i(x_0)) \to p(x^*)$. At the same time, the Jacobian vanishes $\left|\partial f^i / \partial x\right| \to 0$ since the optimizer is a contractive map, implying that their product also vanishes. When the kernel goes to minus infinity ($i \to -\infty$) the density vanishes $p(f^i(x_0)) \to 0$ since the optimizer f minimizes the density going backward. The convergence of this part depends on the special choice of f and how fast the Jacobian of f^{-1} expand the space.

The convergence of weights allows us to truncate the infinite trajectory in both directions. Thus, we can iterate forward and backward from the initial state until the weights ω_i become negligible. The bias introduced by the truncation can be made arbitrarily small by choosing the threshold value. In practice, we iterate while the ratio (the maximum weight)/(the current weight) is less than 10^3 (see Algorithm 3). Note that, due to its discrete nature, the optimization algorithm could be non-monotonic in ω_i. This can be alleviated by choosing smaller step sizes and setting higher threshold W, what might negatively affect the performance.

Several hyperparameter settings related to the optimization dynamics are possible in this algorithm. To avoid expensive evaluations of the Jacobian, similar to HMC, we consider optimization schemes in the joint space of states x and momenta v. It is then easy to see that, for instance, the Jacobian of SGD with momentum is just a constant. In practice, we use the Leap-Frog integrator from (França et al., 2020), which simulates Hamiltonian dynamics with friction. For details, see Appendix B.3.

5 Empirical study[1]

To study the Algorithms 2 and 3, we consider Hamiltonian dynamics for the deterministic function f. Then the joint density $p(x, v, d)$ is the fully-factorized distribution $p(x, v, d) = p(x)\mathcal{N}(v \mid 0, \mathbb{1})p(d)$, where $p(d) = \text{Uniform}\{0, \dots, T-1\}$. For Algorithm 2, the deterministic transition f is the Leap-Frog integrator. For Algorithm 3, we choose the Leap-Frog as well, but with the friction component. We compare both algorithms to HMC and the recycled HMC (Nishimura et al., 2020). Recycled HMC simulates the whole trajectory using the Leapfrog integrator and then, for each point $f^i(x, v)$ of the trajectory, decides whether to collect the sample or not.

To tune the hyperparameters for all algorithms we use the ChEES criterion (Hoffman et al., 2021). During the initial period of adaptation, this criterion optimizes the maximum trajectory length $T_{\max}$ for the HMC with jitter (trajectory length at each iteration is sampled $\sim \text{Uniform}(0, T_{\max})$). To set the stepsize of HMC we follow the common practice of keeping the acceptance rate around 0.65 as suggested in (Beskos et al., 2013). We set this stepsize via double averaging as proposed in (Hoffman & Gelman, 2014) and considered in ChEES-HMC. Note that this choice of hyperparameters is designed especially for HMC and doesn't generalize to other algorithms. We leave the study of adaptation procedures for our algorithms as a future work. For Algorithm 3 we don't need to set the trajectory length, but we use the step size yielded at the adaptation step of ChEES-HMC. The crucial hyperparameter for this algorithm is the friction coefficient β, which we set to $\sqrt[n]{0.8}$, where n is the number of dimensions of the target density, thus setting the contraction rate to 0.64.

For the comparison, we take several target distributions: Banana (2-D), ill-conditioned Gaussian (50-D), the posterior distribution of the Bayesian logistic regression (25-D), and the posterior distribution of the

[1] we provide the code reproducing all experiments at https://github.com/necludov/oMCMC

Table 1: Performance of the algorithms as measured by the Effective Sample Size (ESS) per gradient evaluation (higher values are better). For each algorithm we subsample 1000 points from the trajectory preserving their order. For each of the 100 independent chains we measure the minimum ESS across dimensions and report the median across chains as well as their standard deviations.

Algorithm	Banana	Gaussian	Logistic Reg	Item-response
ChEES-HMC	7.83e–5±6.95e–5	**1.51e–06±6.14e–7**	1.52e–5±6.90e–6	**2.97e–6±1.02e–6**
Recycled-HMC	7.76e–5±9.86e–5	1.43e–6±7.37e–7	2.08e–5±9.28e–6	2.63e–6±8.96e–7
Orbital-HMC	7.71e–5±8.68e–5	1.08e–6±7.30e–7	3.09e–5±1.30e–5	1.73e–6±7.21e–7
Opt-HMC	**2.23e–4±2.41e–4**	2.04e–7±8.90e–8	**3.66e–5±1.82e–5**	2.68e–6±1.53e–6

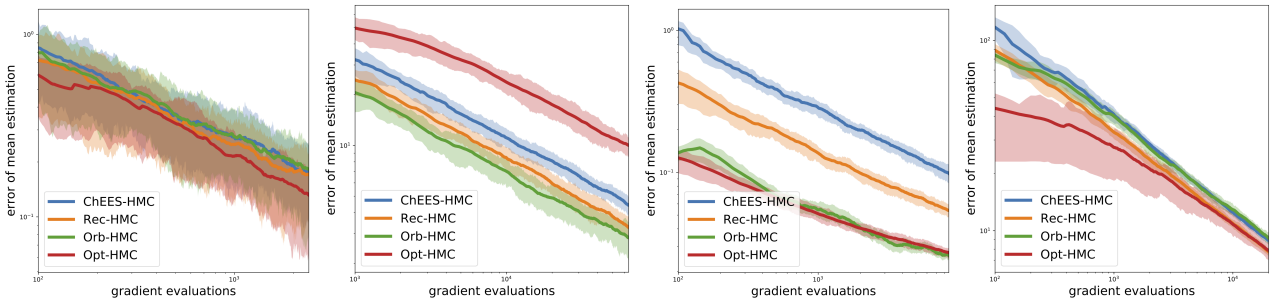

Figure 2: From left to right: the error of mean estimation on Banana, ill-conditioned Gaussian, logistic regression, Item-Response model. Every solid line depicts the mean of the absolute error averaged across 100 independent chains. The shaded area lies between 0.25 and 0.75 quantiles of the error. Orb-HMC corresponds to Algorithm 2 with Hamiltonian dynamics. Opt-HMC corresponds to Algorithm 3 with Hamiltonian with friction.

Item-Response model (501-D) (see description in Appendix B.4). For all experiments, we use 1000 adaptation iterations of ChEES-HMC then followed by 1000 sampling iterations of ChEES-HMC. Then we fix the computational budget in terms of density and gradient evaluations and run all algorithms using approximately the same computational budget.

In Figure 2, we compare the errors in the estimation of the mean of the target distribution as a function of the number of gradient evaluations (which we take as a hardware-agnostic estimation of computation efforts). Algorithm 3 (Opt-HMC) provides the best estimate for the mean value for all distributions except for the ill-conditioned Gaussian (where Orbital-HMC demonstrates the fastest convergence). We can understand the low performance of Opt-HMC for the ill-conditioned Gaussian because of its inability to traverse long distances due to the introduced friction. This property is crucial here since the variances along all dimensions scale logarithmically from 10^{-2} to 10^{2} forcing the chain to keep a small step size. Note that the ChEES criterion was specifically developed for HMC to solve such problems. Another downside of Opt-HMC is that it relatively poorly estimates the variance of the target (we provide corresponding plots in Appendix B.3).

Given the same computational budget, all the algorithms produce different amount of samples. However, the evaluation of statistics of interest could be expensive. Therefore, it is important for algorithms to output a limited amount of low-correlated samples. We validate this property by subsampling the states from the trajectory to yield the same amount of samples as HMC (10^3 samples). We evaluate the Effective Sample Size (ESS) and report the ESS per gradient evaluation (including the adaptation cost into the budget of all algorithms). The results are provided in Table 1. Our algorithms perform relatively poorly on the Gaussian, but perform comparably or better on other distributions. For the Item-Response model, Opt-HMC demonstrates the fastest convergence to the mean, but ChEES-HMC outperforms it in terms of ESS. Thus, the main advantage of the proposed algorithms comes when we are able to use all of the collected samples for the estimation.

6 CONCLUSION

In this paper, we have developed two new practical MCMC algorithms based on iterative deterministic maps. We believe that our orbital MCMC framework opens the door to cross-fertilization between (possibly

chaotic) dynamical systems theory, optimization, and MCMC algorithm design. In Appendix B.5, we discuss possible applications of the oMCMC framework in the context of neural models.

The proposed acceptance rules are opposite to the conventional Metropolis-Hastings test. Namely, in Sections 3 and 4, we apply the kernel an infinite amount of times and accept the whole orbit based on the derived limit. However, we think that the most efficient acceptance strategy would be to minimize the rejection probability: the probability of staying at the same point. Although, it would require either much more computations (if performed straightforwardly) or complex analysis estimating the minimizing time.

Recently, Thin et al. (2021) proposed an importance sampling scheme for the orbits of deterministic transforms. Compared to our methods, it doesn't require periodicity of the orbits or weights vanishing for the infinite orbits. However, it requires additional iterations for the unbiased evaluation of the importance weights. This can be considered as a complementary approach to the techniques considered in the current paper.

Acknowledgements

We thank Evgenii Egorov for providing valuable comments on early versions of the paper. We also thank numerous anonymous reviewers and area-chairs for assessing the paper and contributing to its quality.

References

Albergo, M. S., Kanwar, G., Racanière, S., Rezende, D. J., Urban, J. M., Boyda, D., Cranmer, K., Hackett, D. C., and Shanahan, P. E. Flow-based sampling for fermionic lattice field theories. *arXiv preprint arXiv:2106.05934*, 2021.

Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

Barker, A. A. Monte carlo calculations of the radial distribution functions for a proton? electron plasma. *Australian Journal of Physics*, 18(2):119–134, 1965.

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J.-M., Stuart, A., et al. Optimal tuning of the hybrid monte carlo algorithm. *Bernoulli*, 19(5A):1501–1534, 2013.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

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

França, G., Sulam, J., Robinson, D. P., and Vidal, R. Conformal symplectic and relativistic optimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124008, 2020.

Hoffman, M., Sountsov, P., Dillon, J. V., Langmore, I., Tran, D., and Vasudevan, S. Neutra-lizing bad geometry in hamiltonian monte carlo using neural transport. *arXiv preprint arXiv:1903.03704*, 2019.

Hoffman, M., Radul, A., and Sountsov, P. An adaptive-mcmc scheme for setting trajectory lengths in hamiltonian monte carlo. In *International Conference on Artificial Intelligence and Statistics*, pp. 3907–3915. PMLR, 2021.

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

Kanwar, G., Albergo, M. S., Boyda, D., Cranmer, K., Hackett, D. C., Racanière, S., Rezende, D. J., and Shanahan, P. E. Equivariant flow-based sampling for lattice gauge theory. *arXiv preprint arXiv:2003.06413*, 2020.

Murray, I. and Elliott, L. T. Driving markov chain monte carlo with a dependent random stream. *arXiv preprint arXiv:1204.3187*, 2012.

Neklyudov, K., Welling, M., Egorov, E., and Vetrov, D. Involutive mcmc: a unifying framework. *arXiv preprint arXiv:2006.16653*, 2020.

Nishimura, A., Dunson, D., et al. Recycling intermediate steps to improve hamiltonian monte carlo. *Bayesian Analysis*, 15(4):1087–1108, 2020.

Peskun, P. H. Optimum monte-carlo sampling using markov chains. *Biometrika*, 60(3):607–612, 1973.

Rezende, D. J. and Mohamed, S. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

Roberts, G. O., Rosenthal, J. S., et al. General state space markov chains and mcmc algorithms. *Probability surveys*, 1:20–71, 2004.

Song, J., Zhao, S., and Ermon, S. A-nice-mc: Adversarial training for mcmc. In *Advances in Neural Information Processing Systems*, pp. 5140–5150, 2017.

Spanbauer, S., Freer, C., and Mansinghka, V. Deep involutive generative models for neural mcmc. *arXiv preprint arXiv:2006.15167*, 2020.

Thin, A., Janati El Idrissi, Y., Le Corff, S., Ollion, C., Moulines, E., Doucet, A., Durmus, A., and Robert, C. Neo: Non equilibrium sampling on the orbits of a deterministic transform. *Advances in Neural Information Processing Systems*, 34, 2021.

Turitsyn, K. S., Chertkov, M., and Vucelja, M. Irreversible monte carlo algorithms for efficient sampling. *Physica D: Nonlinear Phenomena*, 240(4-5):410–414, 2011.

Supplementary Material: Orbital MCMC

A PROOFS

A.1 Orbital test

Consider the transition kernel

$$k(x' \mid x) = \delta(x' - f^m(x))g(x) + \delta(x' - x)(1 - g(x)), \;\; g(x) = \inf_{k\in\mathbb{Z}} \left\{ \frac{p(f^k(x))}{p(x)} \left| \frac{\partial f^k}{\partial x} \right| \right\} \tag{29}$$

Substituting this kernel into the stationary condition

$$\int dx \; k(x' \mid x)p(x) = p(x'), \tag{30}$$

we obtain

$$\int dx \; \delta(x' - f^m(x)) \inf_{k\in\mathbb{Z}} \left\{ p(f^k(x)) \left| \frac{\partial f^k}{\partial x} \right| \right\} + \int dx \; \delta(x' - x) \left(p(x) - \inf_{k\in\mathbb{Z}} \left\{ p(f^k(x)) \left| \frac{\partial f^k}{\partial x} \right| \right\} \right) = p(x'). \tag{31}$$

This implies

$$\inf_{k\in\mathbb{Z}} \left\{ p(f^{k-m}(x')) \left| \frac{\partial f^k}{\partial x} \right|_{x=f^{-m}(x')} \right\} \left| \frac{\partial f^{-m}}{\partial x'} \right| = \inf_{k\in\mathbb{Z}} \left\{ p(f^k(x')) \left| \frac{\partial f^k}{\partial x'} \right| \right\} \tag{32}$$

$$\inf_{k\in\mathbb{Z}} \left\{ p(f^{k-m}(x')) \left| \frac{\partial f^{k-m}}{\partial x'} \right| \right\} = \inf_{k\in\mathbb{Z}} \left\{ p(f^k(x')) \left| \frac{\partial f^k}{\partial x'} \right| \right\}. \tag{33}$$

The last equation holds since the shift of the index does not affect the set of lower bounds.

A.2 Returning orbit

We call orbit $\mathrm{orb}(x_0)$ *returning* if for any $x \in \mathrm{orb}(x_0)$ we have $\inf_{k>0} \left\| f^k(x) - x \right\| = 0$. We also can think of this type of orbits as dense orbits on itself. Note that returning orbits is a wider class than periodic orbits. For instance, all of the orbits of $f(x) = (x + a) \bmod 1$ in $[0, 1]$ are returning for any irrational a, but not periodic.

Proposition. *(Returning orbit)*
For continuous target density $p(x)$, consider the kernel (11) *from Theorem 1. If the orbit* $\mathrm{orb}(x)$ *is returning and f preserves a measure with continuous density, then the acceptance test can be written as*

$$g(x) = \inf_{k\in\mathbb{Z}} \left\{ \frac{p(f^k(x))}{p(x)} \left| \frac{\partial f^k}{\partial x} \right| \right\} = \inf_{k\geq 0} \left\{ \frac{p(f^k(x))}{p(x)} \left| \frac{\partial f^k}{\partial x} \right| \right\}. \tag{34}$$

Here we provide the proof for the case of returning orbits ($\forall x \in \mathrm{orb} \inf_{k>0} \left\| f^k(x) - x \right\| = 0$) that are not periodic, i.e. there is no such $k > 0$ that $f^k(x) = x$. By the assumption the deterministic map f preserves some measure with continuous density $q(x)$ on the orbit $\mathrm{orb}(x)$:

$$q(x) = \left| \frac{\partial f^k}{\partial x} \right| q(f^k(x)). \tag{35}$$

Then, we can rewrite the acceptance test from Theorem 1 as

$$g(x) = \inf_{k\in\mathbb{Z}} \left\{ \frac{p(f^k(x))}{p(x)} \left| \frac{\partial f^k}{\partial x} \right| \right\} = \frac{q(x)}{p(x)} \inf_{k\in\mathbb{Z}} \left\{ \frac{p(f^k(x))}{q(f^k(x))} \right\} \tag{36}$$

Using the "returning" property of the orbit $\text{orb}(x)$ we can prove

$$\frac{p(f^n(x))}{q(f^n(x))} \geq \inf_{k\geq 0}\left\{\frac{p(f^k(x))}{q(f^k(x))}\right\} \quad \forall n \in \mathbb{Z}. \tag{37}$$

Indeed, for $n \geq 0$, this holds by the definition of infimum. For $n < 0$, we consider $x^* = f^n(x)$ as a starting point, and use the definition of the returning orbit: $\inf_{k>0} \left\|f^k(x^*) - x^*\right\| = 0$. We define the subsequence of the sequence $\{f^k(x^*)\}$ that converges to x^* as follows. Let k_i be the first number greater than zero such that $\left\|f^{k_i}(x^*) - x^*\right\| \leq 1/i$, then $\lim_{i\to\infty} f^{k_i}(x^*) = x^*$. This infinite subsequence exists since $\inf_{k>0} \left\|f^k(x^*) - x^*\right\| = 0$ but there is no such $k > 0$ that $f^k(x) = x$. By continuity of p and q,

$$\lim_{i\to\infty} \frac{p(f^{k_i}(x^*))}{q(f^{k_i}(x^*))} = \frac{p(x^*)}{q(x^*)} = \frac{p(f^n(x))}{q(f^n(x))}. \tag{38}$$

By the definition of the infimum,

$$\frac{p(f^{k_i}(x^*))}{q(f^{k_i}(x^*))} \geq \inf_{k\geq 0}\left\{\frac{p(f^k(x))}{q(f^k(x))}\right\} \quad k_i > -n, \quad \Longrightarrow \quad \frac{p(f^n(x))}{q(f^n(x))} \geq \inf_{k\geq 0}\left\{\frac{p(f^k(x))}{q(f^k(x))}\right\}. \tag{39}$$

Since

$$\frac{p(f^n(x))}{q(f^n(x))} \geq \inf_{k\geq 0}\left\{\frac{p(f^k(x))}{q(f^k(x))}\right\} \quad \forall n \in \mathbb{Z}, \tag{40}$$

then the infimum over the positive k is a lower bound for $p(f^n(x))/q(f^n(x)) \quad \forall n$. Since the infimum is the maximal lower bound, we have

$$\inf_{k\in\mathbb{Z}}\left\{\frac{p(f^k(x))}{q(f^k(x))}\right\} \geq \inf_{k\geq 0}\left\{\frac{p(f^k(x))}{q(f^k(x))}\right\}. \tag{41}$$

At the same time the infimum over all integers is less than the infimum over positive integers. Thus, we have

$$\inf_{k\in\mathbb{Z}}\left\{\frac{p(f^k(x))}{q(f^k(x))}\right\} = \inf_{k\geq 0}\left\{\frac{p(f^k(x))}{q(f^k(x))}\right\}. \tag{42}$$

A.3 Reversibility

A.3.1 Single kernel

Consider the kernel

$$k_m(x' \,|\, x) = \delta(x' - f^m(x))g(x) + \delta(x' - x)(1 - g(x)), \tag{43}$$

and assume that it satisfies the fixed point equation ($\int dx\; k_m(x' \,|\, x)p(x) = p(x')$) what reduces to the system of equations at each point x':

$$g(x')p(x') = g(f^k(x'))p(f^k(x'))\left|\frac{\partial f^k}{\partial x}\right|_{x=x'} \quad \forall k \in \mathbb{Z}. \tag{44}$$

Since it satisfies the fixed point equation, we can define the reverse kernel

$$r_m(x' \,|\, x) = k_m(x \,|\, x')\frac{p(x')}{p(x)} = \delta(x - f^m(x'))\frac{p(x')}{p(x)}g(x') + \delta(x - x')\left(\frac{p(x')}{p(x)} - \frac{p(x')}{p(x)}g(x')\right). \tag{45}$$

Integrating around the point $x' = x$, we obtain (assuming that $f^{-m}(x) \neq x$)

$$\int_{A(x)} dx'\; r_m(x' \,|\, x) = (1 - g(x)) = \int_{A(x)} dx'\; k_m(x' \,|\, x). \tag{46}$$

Integrating around the point $x' = f^{-m}(x)$, we obtain

$$\int_{A(f^{-m}(x))} dx'\ r_m(x' \,|\, x) = \int_{A(f^{-m}(x))} dx'\ \delta(x - f^m(x'))g(x')\frac{p(x')}{p(x)} = (x' = f^{-m}(y)) \tag{47}$$

$$= \int_{f^m(A(f^{-m}(x)))} dy\ \delta(x-y)g(f^{-m}(y))\frac{p(f^{-m}(y))}{p(x)}\left|\frac{\partial f^{-m}}{\partial y}\right| \tag{48}$$

$$= g(f^{-m}(x))\frac{p(f^{-m}(x))}{p(x)}\left|\frac{\partial f^{-m}}{\partial x}\right| = g(x) \tag{49}$$

And for the forward kernel

$$\int_{A(f^{-m}(x))} dx'\ k_m(x' \,|\, x) = \int_{A(f^{-m}(x))} dx'\delta(x' - f^m(x))g(x) \tag{50}$$

Since m is a fixed number, we can choose the radius of $A(f^{-m}(x))$ small enough that $f^m(x)$ does not lie in $A(f^{-m}(x))$. However, if $f^m(x) = f^{-m}(x)$, we have

$$\int_{A(f^{-m}(x))} dx'\ k_m(x' \,|\, x) = g(x) = \int_{A(f^{-m}(x))} dx'\ r_m(x' \,|\, x). \tag{51}$$

Hence, the kernel $k_m(x' \,|\, x)$ is reversible if and only if $f^{-m}(x') = f^m(x')$.

A.3.2 Linear combination

Consider the kernel

$$k_m(x' \,|\, x) = \delta(x' - f^m(x))g_m(x) + \delta(x' - x)(1 - g_m(x)), \tag{52}$$

that preserves target measure $p(x)$ for any integer m. Then the linear combination

$$k(x' \,|\, x) = \sum_{m\in\mathbb{Z}} w_m k_m(x' \,|\, x), \quad \sum_{m\in\mathbb{Z}} w_m = 1, \quad w_m \geq 0 \tag{53}$$

also preserves the target measure. Hence, we can define the reverse kernel as

$$r(x' \,|\, x) = k(x \,|\, x')\frac{p(x')}{p(x)} = \sum_{m\in\mathbb{Z}} w_m r_m(x' \,|\, x), \tag{54}$$

where r_m is the reverse kernel of k_m as we derived in Appendix A.3.1:

$$r_m(x' \,|\, x) = k_m(x \,|\, x')\frac{p(x')}{p(x)} = \delta(x - f^m(x'))\frac{p(x')}{p(x)}g_m(x') + \delta(x - x')\left(\frac{p(x')}{p(x)} - \frac{p(x')}{p(x)}g_m(x')\right). \tag{55}$$

Further, we assume that all the points of the orbits are distinct, i.e. orbit is neither periodic nor stationary. Integrating the kernel $r(x' \,|\, x)$ around the point $x' = x$, we get

$$\int_{A(x)} dx'\ r(x' \,|\, x) = \sum_{m\in\mathbb{Z}} w_m \int_{A(x)} dx'\ r_m(x' \,|\, x) = \sum_{m\in\mathbb{Z}} w_m(1 - g_m(x)) = \int_{A(x)} dx'\ k(x' \,|\, x). \tag{56}$$

Integrating around the point $x' = f^{-m}(x)$, we obtain

$$\int_{A(f^{-m}(x))} dx'\ r(x' \,|\, x) = \int_{A(f^{-m}(x))} dx'\ w_m\delta(x - f^m(x'))g_m(x')\frac{p(x')}{p(x)} = (x' = f^{-m}(y)) \tag{57}$$

$$= \int_{f^m(A(f^{-m}(x)))} dy\ w_m\delta(x-y)g_m(f^{-m}(y))\frac{p(f^{-m}(y))}{p(x)}\left|\frac{\partial f^{-m}}{\partial y}\right| \tag{58}$$

$$= w_m g_m(f^{-m}(x))\frac{p(f^{-m}(x))}{p(x)}\left|\frac{\partial f^{-m}}{\partial x}\right| = w_m g_m(x) \tag{59}$$

And for the forward kernel

$$\int_{A(f^{-m}(x))} dx'\ k(x' \,|\, x) = \int_{A(f^{-m}(x))} dx'\ w_{-m} k_{-m}(x' \,|\, x) = \tag{60}$$

$$= \int_{A(f^{-m}(x))} dx' w_{-m} \delta(x' - f^{-m}(x)) g_{-m}(x) = w_{-m} g_{-m}(x) \tag{61}$$

In all these derivations we integrated the kernels around a single point assuming that the area $A(x)$ around the point x can be chosen so small that includes x and does not include other points from the orbit. However, this is not the case for returning orbits. By definition, the returning orbit has a subsequence on the orbit that converges to the starting point. Nevertheless, we can choose such small area $A(x)$ that the point weights of the subsequence inside the area $A(x)$ go to zero since all the weights are supposed to be positive and sum up to 1.

Thus, we see that for aperiodic orbits of f the linear combination of kernels $k(x' \,|\, x) = \sum_{m \in \mathbb{Z}} w_m k_m(x' \,|\, x)$ is reversible if and only if $w_{-m} g_{-m}(x) = w_m g_m(x),\ \ \forall m \in \mathbb{Z}$. The same criterion can be formulated for the periodic orbits as well if we assume that the only positive weights are $w_1, \dots, w_{T-1}$, where T is the period of the orbit. Then the linear combination is reversible if and only if $w_{T-m} g_{-m}(x) = w_m g_m(x),\ \ \forall m \in [1, T-1]$.

A.4 Mean ergodic theorem for the escaping orbital kernel

We analyse the kernel

$$k(x' \,|\, x) = \delta(x' - f(x)) g(x) + \delta(x' - x)(1 - g(x)) \tag{62}$$

that preserves target measure $p(x)$, and $g(x) > 0$. Applying this kernel to the delta function $\delta(x - x_0)$ we obtain

$$\int dx\ k(x' \,|\, x) \delta(x - x_0) = (1 - g(x_0)) \delta(x' - x_0) + g(x_0) \delta(x' - f(x_0)). \tag{63}$$

Thus, the kernel leaves a part of mass $(1 - g(x_0))$ at the initial point and propagates the other part $g(x_0)$ further in orbit. Let the initial distribution be the delta-function $p_0(x) = \delta(x - x_0)$. Iteratively applying the kernel $k(x' \,|\, x)$, at time $t+1$, we obtain a chain of weighted delta-functions along the orbit $\text{orb}(x_0)$ of f:

$$\int dx\ k(x' \,|\, x) p_t(x) = p_{t+1}(x') = \sum_{i=0}^{t+1} \omega_i^{t+1} \delta(x' - f^i(x_0)). \tag{64}$$

Denoting the points on the orbit as $x_i = f^i(x_0)$, we can write the recurrence relation for the weights of the delta-function at x_i as

$$\omega_i^{t+1} = (1 - g(x_i)) \omega_i^t + g(x_{i-1}) \omega_{i-1}^t, \ \ t > 0,\ i > 0. \tag{65}$$

For periodic orbits with period T, we consider the lower index of ω_i^{t+1} by modulo T. Further, denoting the sum of the weights over time as

$$S_i^t = \sum_{t'=0}^{t} \omega_i^{t'}, \tag{66}$$

we obtain the following recurrence relation by summation of (65).

$$S_i^{t+1} - \omega_i^0 = (1 - g(x_i)) S_i^t + g(x_{i-1}) S_{i-1}^t \tag{67}$$

Then we denote the solution of this relation as $S_i^t = \alpha_i^t \widehat{S}_i^t$, where $\widehat{S}_i^t$ is the solution of the corresponding homogeneous relation, and α_i^t is the constant variation.

$$\alpha_i^{t+1} \widehat{S}_i^{t+1} = (1 - g(x_i)) \alpha_i^t \widehat{S}_i^t + g(x_{i-1}) S_{i-1}^t + \omega_i^0 \tag{68}$$

$$\alpha_i^{t+1} (1 - g(x_i)) \widehat{S}_i^t = (1 - g(x_i)) \alpha_i^t \widehat{S}_i^t + g(x_{i-1}) S_{i-1}^t + \omega_i^0 \tag{69}$$

$$\alpha_i^{t+1} = \alpha_i^t + \frac{g(x_{i-1}) S_{i-1}^t + \omega_i^0}{(1 - g(x_i)) \widehat{S}_i^t} \tag{70}$$

$$\alpha_i^t = \alpha_i^0 + \sum_{t'=0}^{t-1} \frac{g(x_{i-1}) S_{i-1}^{t'} + \omega_i^0}{(1 - g(x_i)) \widehat{S}_i^{t'}}, \ \ t > 0 \tag{71}$$

Together with the solution $\widehat{S}_i^t = (1 - g(x_i))^t \widehat{S}_i^0$ of the homogeneous relation, we have

$$S_i^t = \left(\alpha_i^0 + \sum_{t'=0}^{t-1} \frac{g(x_{i-1}) S_{i-1}^{t'} + \omega_i^0}{(1 - g(x_i))^{t'+1} S_i^0}\right)(1 - g(x_i))^t \widehat{S}_i^0 \tag{72}$$

$$S_i^t = (1 - g(x_i))^t S_i^0 + \sum_{t'=0}^{t-1} \frac{g(x_{i-1}) S_{i-1}^{t'} + \omega_i^0}{(1 - g(x_i))^{t'-(t-1)}} \tag{73}$$

The first term of the solution goes to zero. Hence, for aperiodic orbits we have

$$\lim_{t\to\infty} S_0^t = \frac{1}{g(x_0)} \tag{74}$$

since $S_{-1}^t = 0$, and $\omega_0^0 = 1$. For $i > 0$, $\omega_i^0 = 0$, and the last sum we can split as

$$\lim_{t\to\infty} S_i^t = \lim_{t\to\infty} \left[(1 - g(x_i))^{t-1} \sum_{t'=0}^{t_1} \frac{g(x_{i-1}) S_{i-1}^{t'}}{(1 - g(x_i))^{t'}} + \sum_{t'=t_1}^{t-1} \frac{g(x_{i-1}) S_{i-1}^{t'}}{(1 - g(x_i))^{t'-(t-1)}}\right]. \tag{75}$$

Taking t_1 large enough, we obtain

$$\lim_{t\to\infty} S_1^t = \lim_{t\to\infty} \sum_{t'=t_1}^{t-1} \frac{g(x_0)/g(x_0)}{(1 - g(x_1))^{t'-(t-1)}} = \lim_{t\to\infty} \sum_{t'=0}^{t-1-t_1} (1 - g(x_1))^{t'} = \frac{1}{g(x_1)}. \tag{76}$$

Applying this recursively, we obtain

$$\lim_{t\to\infty} S_i^t = \lim_{t\to\infty} \sum_{t'=0}^{t} \omega_i^{t'} = \frac{1}{g(x_i)}, \;\; i \geq 0. \tag{77}$$

Note that the limiting behaviour of the chain on the aperiodic orbit depends on the initial conditions $\omega_0^0, \ldots, \omega_i^0, \ldots$, and each individual weight goes to zero with time: $\lim_{t\to\infty} \omega_i^t = 0$. Thus, the total mass drifts away from the initial point down the orbit.

The chain behaves differently in the case of the periodic orbit. Since the number of points on the orbit is finite, the total mass S_i^t goes to infinity with number of iterations t (if $g(x_i) < 1$). However, the average mass over iterations remains constant

$$\frac{1}{t} \sum_{i=0}^{T-1} S_i^t = 1, \;\; t \geq 0, \tag{78}$$

where T is the period of orbit $\mathrm{orb}(x_0)$. Thus, we provide the

$$\lim_{t\to\infty} \frac{1}{t} S_i^t = \lim_{t\to\infty} \frac{1}{t} \sum_{t'=0}^{t-1} \frac{g(x_{i-1}) S_{i-1}^{t'} + \omega_i^0}{(1 - g(x_i))^{t'-(t-1)}} = \tag{79}$$

$$= \lim_{t\to\infty} \left[g(x_{i-1}) \sum_{t'=0}^{t-1} \frac{S_{i-1}^{t'}}{t} (1 - g(x_i))^{(t-1)-t'} + \frac{\omega_i^0}{t} \sum_{t'=0}^{t-1} (1 - g(x_i))^{t'}\right] \tag{80}$$

The second series goes to zero when $t \to \infty$. The first series converges since $S_{i-1}^{t'}/t \leq 1, \forall t'$. Denoting the limit of S_i^t/t as S_i and putting it into the original relation (67), we get

$$S_i = (1 - g(x_i)) S_i + g(x_{i-1}) S_{i-1} \implies S_i = \frac{g(x_{i-1})}{g(x_i)} S_{i-1}. \tag{81}$$

Finally, taking the limit in (78), we have

$$\sum_{i=0}^{T-1} S_i = 1 \implies \sum_{i=0}^{T-1} S_0 \frac{g(x_0)}{g(x_i)} = 1 \implies S_0 = \frac{\frac{1}{g(x_0)}}{\sum_{i=0}^{T-1} \frac{1}{g(x_i)}} \implies S_j = \frac{\frac{1}{g(x_j)}}{\sum_{i=0}^{T-1} \frac{1}{g(x_i)}}. \tag{82}$$

By the assumption, the kernel $k(x' \,|\, x)$ preserves the target distribution $p(x)$. Hence, for the test $g(x)$, we have

$$\frac{g(x_j)}{g(x_i)} = \frac{p(x_i)\left|\frac{\partial f^i}{\partial x}\right|_{x=x_0}}{p(x_j)\left|\frac{\partial f^j}{\partial x}\right|_{x=x_0}}. \tag{83}$$

Thus, for periodic orbits we have

$$\lim_{t\to\infty}\frac{1}{t}S_i^t = \lim_{t\to\infty}\frac{1}{t}\sum_{t'=0}^{t}\omega_i^{t'} = \frac{p(f^i(x_0))\left|\frac{\partial f^i}{\partial x}\right|_{x=x_0}}{\sum_{j=1}^{T}p(f^j(x_0))\left|\frac{\partial f^j}{\partial x}\right|_{x=x_0}}. \tag{84}$$

Note that the limit does not depend on the initial values of the weights $\omega_0^0, \dots, \omega_{T-1}^0$.

A.5 Average escape time

Consider the kernel

$$k(x' \,|\, x) = \delta(x' - f(x))g(x) + \delta(x' - x)(1 - g(x)) \tag{85}$$

that preserves target measure $p(x)$, and $g(x) > 0$. Here we treat it stochastically assuming that we accept $f(x)$ with the acceptance probability $g(x)$ and stay at the same point x with probability $(1 - g(x))$. Assuming that we start at some point x_0, we define the escape time t_n as the number of iterations required to leave x_{n-1}, or equivalently the time of the first acceptance of $x_n = f^n(x_0)$. Then the expectation of t_n is

$$\mathbb{E}t_n = \sum_{t_0,\dots,t_{n-1}}^{\infty}\left(\sum_{i=0}^{n-1}(t_i+1)\right)\prod_{i=0}^{n-1}g(x_i)(1-g(x_i))^{t_i} = \tag{86}$$

$$= \sum_{t_1,\dots,t_{n-1}}^{\infty}\left(\sum_{i=1}^{n-1}(t_i+1)\right)\prod_{i=0}^{n-1}g(x_i)(1-g(x_i))^{t_i} + \tag{87}$$

$$+ \sum_{t_1,\dots,t_{n-1}}^{\infty}\sum_{t_0=0}^{\infty}(t_0+1)\prod_{i=0}^{n-1}g(x_i)(1-g(x_i))^{t_i} \tag{88}$$

The sum over t_0 in the last term is

$$\sum_{t_0=0}^{\infty}(t_0+1)g(x_0)(1-g(x_0))^{t_0} = 1 + \underbrace{\sum_{t_0=1}^{\infty}t_0 g(x_0)(1-g(x_0))^{t_0}}_{S}, \tag{89}$$

where the last sum can be calculated via the index shifting trick as follows.

$$S = \sum_{t_0=1}^{\infty}t_0 g(x_0)(1-g(x_0))^{t_0} = g(x_0)(1-g(x_0)) + \sum_{t_0=2}^{\infty}t_0 g(x_0)(1-g(x_0))^{t_0} = \tag{90}$$

$$= g(x_0)(1-g(x_0)) + \sum_{t_0=1}^{\infty}(t_0+1)g(x_0)(1-g(x_0))^{t_0+1} = \tag{91}$$

$$= g(x_0)(1-g(x_0)) + (1-g(x_0))\underbrace{\sum_{t_0=1}^{\infty}t_0 g(x_0)(1-g(x_0))^{t_0}}_{S} + (1-g(x_0))^2 \tag{92}$$

Hence,

$$S = (1-g(x_0)) + \frac{1}{g(x_0)} - 2 + g(x_0) = \frac{1}{g(x_0)} - 1. \tag{93}$$

Substituting all the derivations back into the formula for expectation, we have

$$\mathbb{E}t_n = \sum_{t_1,\dots,t_{n-1}}^{\infty} \left(\sum_{i=1}^{n-1} (t_i + 1) \right) \prod_{i=0}^{n-1} g(x_i)(1 - g(x_i))^{t_i} + \tag{94}$$

$$+ \sum_{t_1,\dots,t_{n-1}}^{\infty} \sum_{t_0=0}^{\infty} (t_0 + 1) \prod_{i=0}^{n-1} g(x_i)(1 - g(x_i))^{t_i} = \tag{95}$$

$$= \sum_{t_1,\dots,t_{n-1}}^{\infty} \left(\sum_{i=1}^{n-1} (t_i + 1) \right) \prod_{i=0}^{n-1} g(x_i)(1 - g(x_i))^{t_i} + \frac{1}{g(x_0)}. \tag{96}$$

Applying this reasoning $(n-1)$ more times, we obtain

$$\mathbb{E}t_n = \sum_{i=0}^{n-1} \frac{1}{g(x_i)}. \tag{97}$$

A.6 Diffusing orbital kernel

We start with the kernel

$$k(x' \,|\, x) = \delta(x' - f(x))g^+(x) + \delta(x' - f^{-1}(x))g^-(x) + \delta(x - x')(1 - g^+(x) - g^-(x)). \tag{98}$$

Putting this kernel into $Kp = p$, we get

$$\begin{aligned} \int dx\ \delta(x' - f(x))g^+(x)p(x) + \int dx\ \delta(x' - f^{-1}(x))g^-(x)p(x) + \\ + p(x')(1 - g^+(x') - g^-(x')) = p(x'), \end{aligned} \tag{99}$$

which yields the following condition for $Kp = p$.

$$g^+(f^{-1}(x'))p(f^{-1}(x'))\left|\frac{\partial f^{-1}}{\partial x'}\right| + g^-(f(x'))p(f(x'))\left|\frac{\partial f}{\partial x'}\right| = g^+(x')p(x') + g^-(x')p(x') \tag{100}$$

One of the solutions here is the linear combination of two escaping orbital kernels, which can be obtained by separately matching terms with g^+ and then the terms with g^-. Another solution, which yields the diffusing orbital kernel, is obtained by matching

$$g^+(f^{-1}(x'))p(f^{-1}(x'))\left|\frac{\partial f^{-1}}{\partial x'}\right| = g^-(x')p(x') \quad \text{and} \quad g^-(f(x'))p(f(x'))\left|\frac{\partial f}{\partial x'}\right| = g^+(x')p(x'). \tag{101}$$

Considering x' as some point from the orbit $\mathrm{orb}(x_0)$, i.e. $x' = f^k(x_0)$, both equations yield the same system

$$\frac{g^+(f^{k-1}(x_0))}{g^-(f^k(x_0))} = \frac{p(f^k(x_0))}{p(f^{k-1}(x_0))} \frac{\left|\partial f^k/\partial x\right|_{x=x_0}}{\left|\partial f^{k-1}/\partial x\right|_{x=x_0}}, \quad \forall k \in \mathbb{Z}. \tag{102}$$

Now we use the same trick as for the escaping orbital kernel assuming that f preserves some measure with the density $q(x)$, i.e. $q(x_0) = q(f^k(x_0))\left|\partial f^k/\partial x\right|_{x=x_0}$. Expressing the Jacobians in terms of the density q, we obtain

$$\frac{g^+(f^{k-1}(x_0))}{g^-(f^k(x_0))} = \frac{p(f^k(x_0))}{p(f^{k-1}(x_0))} \frac{q(f^{k-1}(x_0))}{q(f^k(x_0))}, \quad \forall k \in \mathbb{Z}. \tag{103}$$

Once again, there are two options to design the test function here. One of the options is to say let $g^+(x) = g^-(x) \propto q(x)/p(x)$, but saying so we will end up with the mixture of escaping orbital kernels. Another option is to consider the tests as follows

$$g^+(x) = \frac{p(f(x))}{q(f(x))}c \quad \text{and} \quad g^-(x) = \frac{p(f^{-1}(x))}{q(f^{-1}(x))}c, \tag{104}$$

where the constant c comes from the fact that the rejection probability is non-negative on the whole orbit $\text{orb}(x_0)$, i.e. $(1 - g^+(x) - g^-(x)) \geq 0$, and we have

$$c\left(\frac{p(f^{k+1}(x_0))}{q(f^{k+1}(x_0))} + \frac{p(f^{k-1}(x_0))}{q(f^{k-1}(x_0))}\right) \leq 1. \tag{105}$$

Thus, we end up with the following value of c

$$c = \inf_{k\in\mathbb{Z}} \left\{ \frac{1}{p(f^{k+1}(x_0))/q(f^{k+1}(x_0)) + p(f^{k-1}(x_0))/q(f^{k-1}(x_0))} \right\}, \tag{106}$$

or with another lower bound, which may be not optimal, but useful in practice:

$$c' = \frac{1}{2} \inf_{k\in\mathbb{Z}} \left\{ \frac{q(f^k(x_0))}{p(f^k(x_0))} \right\} \leq c. \tag{107}$$

Putting these constants back into the tests (104), we get

$$g^+(x) = p(f(x))\left|\frac{\partial f}{\partial x}\right| c(x), \quad g^-(x) = p(f^{-1}(x))\left|\frac{\partial f^{-1}}{\partial x}\right| c(x), \tag{108}$$

where $c(x)$ may be chosen as

$$c'(x) = \frac{1}{2} \inf_{k\in\mathbb{Z}} \left\{ \frac{1}{p(f^k(x))\left|\frac{\partial f^k}{\partial x}\right|} \right\}, \quad \text{or} \quad c(x) = \inf_{k\in\mathbb{Z}} \left\{ \frac{1}{p(f^{k+1}(x))\left|\frac{\partial f^{k+1}}{\partial x}\right| + p(f^{k-1}(x))\left|\frac{\partial f^{k-1}}{\partial x}\right|} \right\}. \tag{109}$$

A.7 Reversibility of the diffusing orbital kernel

Consider the diffusing orbital kernel

$$k(x' \,|\, x) = \delta(x' - f(x))g^+(x) + \delta(x' - f^{-1}(x))g^-(x) + \delta(x' - x)(1 - g^+(x) - g^-(x)), \tag{110}$$

that preserves the target measure ($\int dx\; k_m(x' \,|\, x)p(x) = p(x')$) what reduces to the system of equations:

$$\frac{g^+(f^{k-1}(x_0))}{g^-(f^k(x_0))} = \frac{p(f^k(x_0))}{p(f^{k-1}(x_0))} \frac{\left|\partial f^k/\partial x\right|_{x=x_0}}{\left|\partial f^{k-1}/\partial x\right|_{x=x_0}}, \quad \forall k \in \mathbb{Z}. \tag{111}$$

Since it satisfies the fixed point equation, we can define the reverse kernel

$$r(x' \,|\, x) = k(x \,|\, x')\frac{p(x')}{p(x)} = \delta(x - f(x'))\frac{p(x')}{p(x)}g^+(x') + \delta(x - f^{-1}(x'))\frac{p(x')}{p(x)}g^-(x') + \tag{112}$$

$$+\delta(x - x')\frac{p(x')}{p(x)}(1 - g^+(x') - g^-(x')). \tag{113}$$

Integrating around the point $x' = x$, we obtain

$$\int_{A(x)} dx'\; r(x' \,|\, x) = (1 - g^+(x) - g^-(x)) = \int_{A(x)} dx'\; k(x' \,|\, x). \tag{114}$$

Integrating around the point $x' = f^{-1}(x)$, we obtain

$$\int_{A(f^{-1}(x))} dx'\; r(x' \,|\, x) = \int_{A(f^{-1}(x))} dx'\; \delta(x - f(x'))g^+(x')\frac{p(x')}{p(x)} = (x' = f^{-1}(y)) \tag{115}$$

$$= \int_{f(A(f^{-1}(x)))} dy\; \delta(x - y)g^+(f^{-1}(y))\frac{p(f^{-1}(y))}{p(x)}\left|\frac{\partial f^{-1}}{\partial y}\right| \tag{116}$$

$$= g^+(f^{-1}(x))\frac{p(f^{-1}(x))}{p(x)}\left|\frac{\partial f^{-1}}{\partial x}\right| = g^-(x) \tag{117}$$

And for the forward kernel, the integral around $x' = f^{-1}(x)$ yields

$$\int_{A(f^{-1}(x))} dx'\ k(x' \,|\, x) = \int_{A(f^{-1}(x))} dx' \delta(x' - f^{-1}(x)) g^-(x) = g^-(x). \tag{118}$$

Thus, we have

$$\int_{A(f^{-1}(x))} dx'\ r(x' \,|\, x) = \int_{A(f^{-1}(x))} dx'\ k(x' \,|\, x) \tag{119}$$

The same reasoning applies for the integration around $x' = f(x)$, hence, we conclude that the diffusing orbital kernel is reversible since $r(x' \,|\, x) = k(x' \,|\, x)$.

A.8 Mean ergodic theorem for the diffusing orbital kernel

We analyse the diffusive orbital kernel

$$k(x' \,|\, x) = \delta(x' - f(x)) g^+(x) + \delta(x' - f^{-1}(x)) g^-(x) + \delta(x - x')(1 - g^+(x) - g^-(x)) \tag{120}$$

that preserves target measure $p(x)$, thus having (from Appendix A.6)

$$g^+(x) = \frac{p(f(x))}{q(f(x))} c \quad \text{and} \quad g^-(x) = \frac{p(f^{-1}(x))}{q(f^{-1}(x))} c. \tag{121}$$

We assume that the map f preserves some density q only to simplify the derivations. One can as well consider the tests derived in Appendix A.6.

Applying this kernel to the delta function $\delta(x - x_0)$ we obtain

$$\begin{aligned} \int dx\ k(x' \,|\, x) \delta(x - x_0) =& g^+(x_0) \delta(x' - f(x_0)) + g^-(x_0) \delta(x' - f(x_0)) + \\ &+ (1 - g^+(x_0) - g^-(x_0)) \delta(x' - x_0). \end{aligned} \tag{122}$$

That is, the kernel leaves a part of mass $(1 - g^+(x_0) - g^-(x_0))$ at the initial point and moves the other part to the adjacent points $f(x_0)$ and $f^{-1}(x_0)$. Thus, starting from the delta-function $p_0(x) = \delta(x - x_0)$, and iteratively applying the kernel $k(x' \,|\, x)$, at time $t + 1$, we obtain a chain of weighted delta-functions along the orbit $\mathrm{orb}(x_0)$ of f:

$$\int dx\ k(x' \,|\, x) p_t(x) = p_{t+1}(x') = \sum_{i=-(t+1)}^{t+1} \omega_i^{t+1} \delta(x' - f^i(x_0)). \tag{123}$$

Denoting the points on the orbit as $x_i = f^i(x_0)$, we can write the recurrence relation for the weights of the delta-function at x_i as

$$\omega_i^{t+1} = (1 - g^+(x_i) - g^-(x_i)) \omega_i^t + g^+(x_{i-1}) \omega_{i-1}^t + g^-(x_{i+1}) \omega_{i+1}^t, \ \ t \geq 0. \tag{124}$$

For periodic orbits with period T, we consider the lower index of ω_i^{t+1} by modulo T. Further, denoting the sum of the weights over time as

$$S_i^t = \sum_{t'=0}^{t-1} \omega_i^{t'}, \tag{125}$$

we obtain the following recurrence relation by summation of (124).

$$S_i^{t+1} = \omega_i^0 + (1 - g^+(x_i) - g^-(x_i)) S_i^t + g^+(x_{i-1}) S_{i-1}^t + g^-(x_{i+1}) S_{i+1}^t \tag{126}$$

Decomposing the solution as $S_i^t = \alpha_i^t \widehat{S}_i^t$, where $\widehat{S}_i^t$ is the solution of the corresponding homogeneous relation, and α_i^t is the constant variation, we get

$$S_i^t = (1 - g^+(x_i) - g^-(x_i))^t S_i^0 + \sum_{t'=0}^{t-1} \frac{g^+(x_{i-1}) S_{i-1}^{t'} + g^-(x_{i+1}) S_{i+1}^{t'} + \omega_i^0}{(1 - g^+(x_i) - g^-(x_i))^{t' - (t-1)}}. \tag{127}$$

Then for the time-average we have

$$\frac{S_i^t}{t} = (1 - g^+(x_i) - g^-(x_i))^t \frac{S_i^0}{t} + \sum_{t'=0}^{t-1} \frac{g^+(x_{i-1})S_{i-1}^{t'}/t + g^-(x_{i+1})S_{i+1}^{t'}/t + \omega_i^0/t}{(1 - g^+(x_i) - g^-(x_i))^{t'-(t-1)}}, \tag{128}$$

which clearly converges when $t \to \infty$ since $\sum_i S_i^t = t$, hence, $S_{i+1}^{t'}/t < 1$, $S_{i-1}^{t'}/t < 1$, and

$$\lim_{t\to\infty} \frac{S_i^t}{t} = \lim_{t\to\infty} \sum_{t'=0}^{t-1} \frac{g^+(x_{i-1})S_{i-1}^{t'}/t + g^-(x_{i+1})S_{i+1}^{t'}/t}{(1 - g^+(x_i) - g^-(x_i))^{t'-(t-1)}} \leq \tag{129}$$

$$\leq \lim_{t\to\infty} \sum_{t'=0}^{t-1} \frac{g^+(x_{i-1}) + g^-(x_{i+1})}{(1 - g^+(x_i) - g^-(x_i))^{t'-(t-1)}}. \tag{130}$$

Thus, the time-average converges since it is dominated by the convergent series. To analyse the limit of the time average, we first simplify the notation a little bit by noting that

$$g^+(x) = g(f(x)) \;\text{ and }\; g^-(x) = g(f^{-1}(x)), \;\text{ where }\; g(x) = \frac{p(x)}{q(x)}c. \tag{131}$$

Then we have $g(x_i) = g^+(x_{i-1}) = g^-(x_{i+1})$, $0 < g(x_i) \leq 1$ and $0 \leq g(x_{i+1}) + g(x_{i-1}) < 1$. Denoting $A_i^t = S_i^t/t$ in (126) we write the recurrence relation for the time average

$$A_i^{t+1} = \frac{\omega_i^0}{t} + (1 - g(x_{i+1}) - g(x_{i-1}))A_i^t + g(x_i)(A_{i-1}^t + A_{i+1}^t). \tag{132}$$

Taking the time t large enough we get rid off the initial condition ω_i^0 and consider the stationary point of this recurrence relation.

$$(g(x_{i+1}) + g(x_{i-1}))A_i = g(x_i)(A_{i-1} + A_{i+1}). \tag{133}$$

Further, we consider the stationary point $A_i \propto g(x_i)$, and use the fact that the sum of the average across all points equals to 1: $\sum_i S_i^t/t = \sum_i A_i^t = 1$, where we take the sum over i for aperiodic orbits $i \in \mathbb{Z}$, and for periodic orbits we take i modulo T. Thus, we have

$$A_i^t = \frac{A_i^t}{\sum_j A_j^t} = \frac{g(x_i)}{\sum_j g(x_j)} = \frac{p(x_i)/q(x_i)}{\sum_j p(x_j)/q(x_j)}. \tag{134}$$

Using that $q(x_j)\left|\partial f^j/\partial x\right| = q(x_i)\left|\partial f^i/\partial x\right|$, we obtain

$$A_i^t = \frac{p(x_i)}{\sum_j p(x_j)(q(x_i)/q(x_j))} = \frac{p(x_i)\left|\partial f^i/\partial x\right|}{\sum_j p(x_j)\left|\partial f^j/\partial x\right|}. \tag{135}$$

For periodic orbits, the time-average always converges to

$$A_i^t = \frac{p(x_i)\left|\partial f^i/\partial x\right|}{\sum_{j=0}^{T-1} p(x_j)\left|\partial f^j/\partial x\right|}. \tag{136}$$

For aperiodic orbits, if the series $\sum_{j=-\infty}^{+\infty} p(x_j)\left|\partial f^j/\partial x\right|$ converges, then the time-average converges to

$$A_i^t = \frac{p(x_i)\left|\partial f^i/\partial x\right|}{\sum_{j=-\infty}^{+\infty} p(x_j)\left|\partial f^j/\partial x\right|}. \tag{137}$$

The uniqueness of the limit follows from the mean-ergodic theorem and the stochastic interpretation of the process. Indeed, in such case the kernel $k(x' \,|\, x)$ is irreducible on the orbit $\text{orb}(x_0)$.

B APPLICATIONS

B.1 Linear combination

Algorithm 4 Linear combination of escaping orbital kernels

input target density $p(x)$, auxiliary density $p(v \,|\, x)$ and a sampler from $p(v \,|\, x)$
input continuous bijection $f(x, v)$, weights of the combination $\{w_m\}_{m=0}^{T-1}$
  initialize x
  for $i = 0 \dots n$ **do**
    sample $v \sim p(v \,|\, x)$
    propose $(x_m, v_m) = f^m(x, v),\ m \in [0, T-1]$
    evaluate $p(x_m, v_m)\big|\frac{\partial f^m}{\partial [x,v]}\big|$ for all $m \in [0, T-1]$
    $g_m = \min_{k \in [0,T-1]}\{\frac{p(f^{mk}(x,v))}{p(x,v)}\big|\frac{\partial f^{mk}}{\partial [x,v]}\big|\}\ \ m \in [0, T-1]$ (can be done in parallel)
    $x_m, w_m \leftarrow \begin{cases} x_m, w_m, & \text{with probability } g_m \\ x, w_m, & \text{with probability } (1 - g_m) \end{cases}$
    samples $\leftarrow$ samples $\cup\ \{(x_m, w_m)\}_{m=0}^{T-1}$
    $x \leftarrow x_j$ with probability w_j
  end for
output samples

In this section we discuss how one can make use of the linear combination of escaping orbital kernels. The main benefit of this procedure is that one can evaluate the tests for all powers of f (see Corollary 1) almost for free. With more details, consider the set of kernels $k_m(x' \,|\, x) = \delta(x' - f^m(x))g_m(x) + \delta(x' - x)(1 - g_m(x)),\ \forall m \in \mathbb{Z}$, where $g_m(x)$ are tests that guarantee $K_m p = p$. Then any linear combination of such kernels:

$$k(x' \,|\, x) = \sum_{m \in \mathbb{Z}} w_m k_m(x' \,|\, x), \quad \sum_{m \in \mathbb{Z}} w_m = 1, \quad w_m \geq 0, \tag{138}$$

clearly admits $Kp = p$. Considering periodic orbits (period T), and taking the tests $g_m(x) = \inf_{k \in \mathbb{Z}}\{p(f^{mk}(x))/p(x)\big|\partial f^{mk}/\partial x\big|\}$ (as in Corollary 1) the evaluation of all tests reduces to the measurement of the density (and the determinant of Jacobian) over whole orbit, and then evaluation of minimums across different subsets of points. Once all the tests are calculated, we can simulate accept/reject steps and then collect all the simulated samples x_m with corresponding weights w_m. The next starting point of the chain can be selected by sampling from the collected samples with probabilities w_m. See Algorithm 4.

The reversibility criterion from Theorem 1 naturally extends to the linear combination as follows.

Proposition 4. *(Reversibility of the linear combination)*
Consider the linear combination of kernels (138)*, where for each kernel we have $K_m p = p$, and $g_m(x) > 0$. Then for the linear combination $k(x' \,|\, x)$ we have $Kp = p$, and it is reversible ($k(x' \,|\, x)p(x) = k(x \,|\, x')p(x')$) if and only if $w_{-m} g_{-m}(x) = w_m g_m(x), \forall m \in \mathbb{Z}$. Note that for the test from Corollary 1 we have $g_m(x) = g_{-m}(x)$.*

For periodic orbit $\mathrm{orb}(x)$ *of f, assume that the only positive weights are $w_1, \dots, w_{T-1}$, where T is the period of* $\mathrm{orb}(x)$*. Then the linear combination is reversible if and only if $w_{T-m} g_{T-m}(x) = w_m g_m(x), \forall m \in \{1, \dots, T-1\}$.*

Proof. See Appendix A.3.2 □

B.2 Importance sampling via the orbital kernel

As we outlined in Section 3, the limiting behaviour of the orbital kernel depends on the orbit it operates on. For infinite orbits, the kernel always moves the probability mass further along the orbit escaping any given point on the orbit since there is no mass flowing backward (the weight of $f^{-1}(x_0)$ is zero from the beginning). Intuitively, we can think that the weights ω_i^t from (15) evolve in time as a wave packet moving on the real line (Fig. 3). With this intuition it becomes evident that the time average at every single point on the orbit converges to zero since the mass always escapes this point. Therefore, if we want to accept several points from the orbit, we need to wait until the kernel "leaves" this set of points and then stop the procedure. To avoid the explicit simulation of the kernel, we provide an approximation for the escape time in the following proposition.

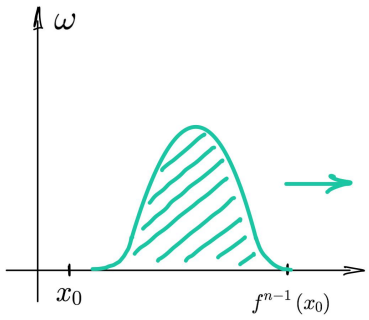

Figure 3: The illustration of the orbital kernel on an infinite orbit.

Proposition 5. *(Average escape time)*
Consider the proper escaping orbital kernel ($Kp = p$, and $g(x) > 0$), and the initial distribution $p_0(x) = \delta(x - x_0)$. The escape time t_n is the number of iterations required to leave the set $\{x_0, f(x_0), \dots, f^{n-1}(x_0)\}$, or equivalently the time of the first acceptance of $f^n(x_0)$. The expectation of t_n is $\mathbb{E}t_n = \sum_{i=0}^{n-1} 1/g(f^i(x_0))$.

Proof. See Appendix A.5. □

Using this proposition, we approximate the time average for points $\{x_0, \dots, f^{n-1}(x_0)\}$ as follows. Once the kernel escaped $f^{n-1}(x_0)$, the sum of each individual weight over time has converged. Moreover, from Theorem 2 we know $\lim_{t\to\infty} \sum_{t'=0}^{t} \omega_i^{t'} = 1/g(f^i(x_0))$. To approximate the time-average we can divide this sum by the average escape time from Proposition 5. Thus, we estimate $p_{n-1}(x) \approx \sum_{i=0}^{n-1} \omega_i \delta(x - f^i(x_0))$, where

$$\omega_i = \frac{1}{\mathbb{E}t_n} \lim_{t\to\infty} \sum_{t'=0}^{t} \omega_i^{t'} = \frac{1/g(f^i(x_0))}{\sum_{j=0}^{n-1} 1/g(f^j(x_0))} = \frac{p(f^i(x_0))\left|\frac{\partial f^i}{\partial x}\right|_{x=x_0}}{\sum_{j=0}^{n-1} p(f^j(x_0))\left|\frac{\partial f^j}{\partial x}\right|_{x=x_0}}. \tag{139}$$

On the last step, the ratio $g(f^i(x_0))/g(f^j(x_0))$ can be found from (8), i.e., using the fact that the kernel K preserves the target density. The resulting formula is tightly related to the self-normalized importance sampling (SNIS) (Andrieu et al., 2003). Indeed, if f preserves some density q on the orbit $\text{orb}(x_0)$, then we can rewrite the Jacobians using this density and put it into (139) as follows.

$$\forall i \quad \left|\frac{\partial f^i}{\partial x}\right|_{x=x_0} = \frac{q(x_0)}{q(f^i(x_0))}; \text{ hence, } \omega_i = \frac{p(f^i(x_0))/q(f^i(x_0))}{\sum_{j=0}^{n-1} p(f^j(x_0))/q(f^j(x_0))} \tag{140}$$

We illustrate the usefulness of the formula (139) with the following example. Consider the target distribution p, and the proposal q. Then one can estimate the mean of the target using the conventional SNIS procedure, which we call stochastic SNIS. Stochastic SNIS operates as follows. It samples from the proposal distribution with density $q(x)$, and then evaluates the weight of sample x_i as $w_i = p(x_i)/q(x_i)$, once all the weights are evaluated it normalizes them as $w_i \leftarrow w_i / \sum_j w_j$. The normalization of the weights is usually done in the setting where the target density is known up to the normalization constant. Another way to collect samples is to consider a deterministic map that preserves q and then re-weight the trajectory of this map (we call it deterministic SNIS). To perform the deterministic SNIS we choose such f that (I) preserves the density of the proposal $q(x)$ (the same as for stochastic SNIS), and (II) has dense aperiodic orbits on the state space. These two goals are achieved by the map described in (Murray & Elliott, 2012).

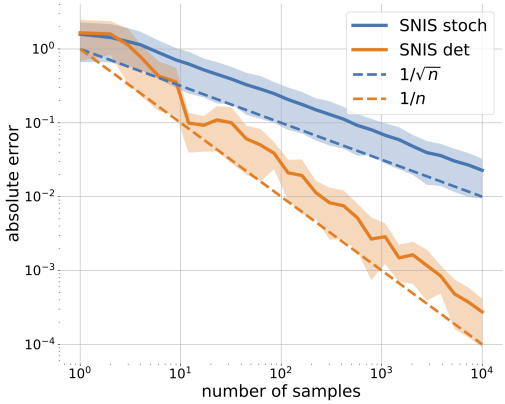

Figure 4: The comparison of deterministic and stochastic SNIS (averaged across 200 independent runs). Solid lines demonstrate the mean error of the estimate, and the shaded area lies between 0.25 and 0.75 quantiles.

Namely, the deterministic map is $f(x) = F^{-1}((F(x) + a) \bmod 1)$, where a is the irrational number (to cover densely the state space) and F is the CDF of q (to preserve q). Another way of thinking about iterations of $f(x)$ is that it firstly sample from Uniform$[0, 1]$ using the Weyl's sequence $u_{i+1} = u_i + a \bmod 1$, and then use the inverse CDF to obtain samples from the desired distribution.

For the target distribution we take the mixture of two Gaussians $p(x) = 0.5 \cdot \mathcal{N}(x \,|\, -2, 1) + 0.5 \cdot \mathcal{N}(x \,|\, 2, 1)$, and for the proposal we take single Gaussian $q(x) = \mathcal{N}(x \,|\, 0, 2)$. For the irrational number a in the deterministic SNIS, we take the float approximation of $\sqrt{2}$. Collecting samples from both procedures we estimate the mean of the target distribution and evaluate the squared error of the estimation. Fig. 4 demonstrates that the deterministic SNIS allows for a more accurate estimate having the same number of samples.

The practical benefit of the formula (139) comes from the fact that we can perform deterministic SNIS even when we don't know the stationary distribution q of the deterministic map.

B.3 Orbital MCMC with Hamiltonian dynamics

Building upon Algorithms 2 and 3, we consider special cases, which use the Hamiltonian dynamics for the deterministic map $f(x, v)$. That is, we consider the joint distribution $p(x, v) = p(x)\mathcal{N}(v \,|\, 0, \mathbb{1})$, and for the map $f(x, v)$ we take the Leapfrog integrator ($f(x, v) = [x', v'']$):

$$\begin{aligned} v' &= v - \frac{\varepsilon}{2}\nabla_x(-\log p(x)), \\ x' &= x + \varepsilon\nabla_v(-\log p(v')), \\ v'' &= v' - \frac{\varepsilon}{2}\nabla_x(-\log p(x')), \end{aligned} \tag{141}$$

where ε is the step-size. Adding the directional variable as proposed in Algorithm 2 we obtain the Orbital-HMC algorithm. Note that the Jacobian of the map is $\left|\partial f(x, v)/\partial[x, v]\right| = 1$.

For the contractive map in Algorithm 3, we consider the same Hamiltonian dynamics but with friction. To add friction, we follow (França et al., 2020), but use the velocity Verlet integrator instead of the coordinate. The resulting integrator is ($f(x, v) = [x', v'']$):

$$\begin{aligned} v' &= \beta\big[v - \frac{\varepsilon}{2}\nabla_x(-\log p(x))\big], \\ x' &= x + \frac{\varepsilon}{2}(\beta^{-1} + \beta)\nabla_v(-\log p(v')), \\ v'' &= \beta\big[v' - \frac{\varepsilon}{2}\nabla_x(-\log p(x'))\big], \end{aligned} \tag{142}$$

where β is the contractive coefficient. Note that $\beta = 1$ yields the standard Leapfrog. The Jacobian of the map is $\left|\partial f(x, v)/\partial[x, v]\right| = \beta^{2n}$, where n is the number of dimensions of the target distribution. For $\beta < 1$, the map becomes contractive with the stationary points at $\nabla_x \log p(x) = 0$. We refer to the resulting algorithm as Opt-HMC due to its optimization properties.

For Opt-HMC, we provide a motivation for the truncation of the weights but leave the rigorous study of dissipative systems beyond the scope of the current work. For the Hamiltonian system with the friction, we have

$$\frac{dx}{dt} = v, \quad \frac{dv}{dt} = -\gamma v - \nabla_x U(x) \quad \Longrightarrow \quad \frac{dH}{dt} = \frac{\partial H}{\partial x}\frac{dx}{dt} + \frac{\partial H}{\partial v}\frac{dv}{dt} = -\gamma\|v\|^2 < 0. \tag{143}$$

Since the Hamiltonian is bounded from below the weights for the forward iterations $\omega(t) = \exp(-H(t) - \gamma n t)$ converges to zero (when $t \to \infty$) exponentially fast after converging to the local minimum of H.

For the backward iterations, the situation is opposite. The Jacobian increases weights with time as $\gamma n t$, while the Hamiltonian increases decreasing the weights. Thus, we have to upper bound the following expression

$$\log \omega(t) = H(0) + \gamma n t - \gamma \int dt \, \|v(t)\|^2. \tag{144}$$

In general, the lower bound asymptotics for $\int dt \, \|v(t)\|^2$ is a difficult question. Based on the equation $dv/dt = \gamma v + \nabla U$, we assume that eventually $\langle v, \nabla U\rangle$ becomes positive and remains positive. Then we can write

$$(\text{if } \langle v, \nabla_x U(x)\rangle \geq 0) \quad \frac{d}{dt}\|v(t)\|^2 = 2\gamma\|v(t)\|^2 + 2\langle v, \nabla_x U(x)\rangle \geq 2\gamma\|v(t)\|^2 \implies \|v\|^2 \geq \exp(2\gamma t). \tag{145}$$

Thus, $\log \omega(t)$ decreases exponentially fast, and the weight $\omega(t)$ decreases as $\exp(-\exp(t))$. In Fig. 1., we provide an illustration of this process in practice. Also, in practice, we compare the algorithms by comparing the errors in statistics estimation, hence, we can argue that the truncation doesn't violate the correctness of sampling.

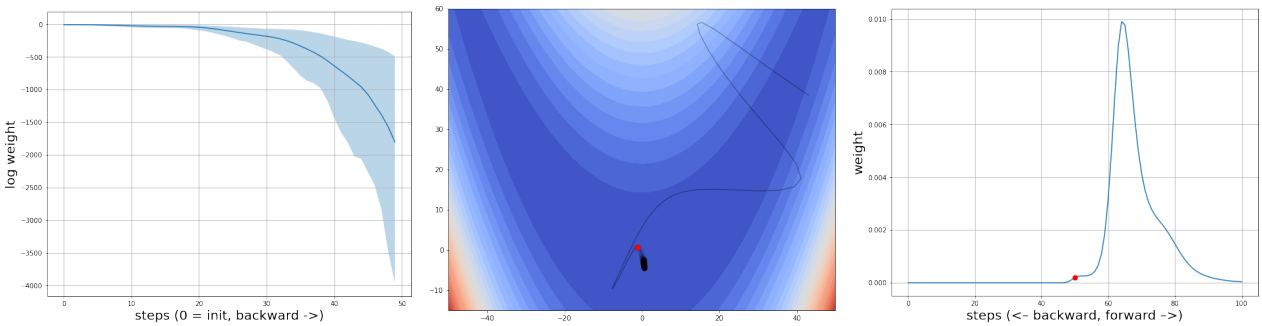

Figure 5: Banana distribution. From left to right: average log-weight for backward iterations with quantiles $(0.1 - 0.9)$; single trajectory example on the energy landscape of Banana; weights of the trajectory points. Initial point is denoted by the red dot.

To tune the hyperparameters for all algorithms we use the ChEES criterion (Hoffman et al., 2021). During the initial period of adaptation, this criterion optimizes the maximum trajectory length $T_{\max}$ for the HMC with jitter (trajectory length at each iteration is sampled $\sim \text{Uniform}(0, T_{\max})$). To set the stepsize of HMC we follow the common practice of keeping the acceptance rate around 0.65 as suggested in (Beskos et al., 2013). We set this stepsize via double averaging as proposed in (Hoffman & Gelman, 2014) and considered in ChEES-HMC. For Opt-HMC, we don't need to set the trajectory length, but we use the step size yielded at the adaptation step of ChEES-HMC. The crucial hyperparameter for this algorithm is the friction coefficient β, which we set to $\sqrt[n]{0.8}$, where n is the number of dimensions of the target density, thus setting the contraction rate to 0.64.

In Figures 6 and 7, we compare the errors in the estimation of the mean and the variance of the target distribution as a function of the number of gradient evaluations (which we take as a hardware-agnostic estimation of computation efforts). Opt-HMC provides the best estimate for the mean value for all distributions except for the ill-conditioned Gaussian (where Orbital-HMC demonstrates the fastest convergence). Another downside of Opt-HMC is that it relatively poorly estimates the variance of the target as provided in Fig. 7, which could also be explained by the introduced contraction of space. Note that Orbital-HMC always performs comparably or better than competitors.

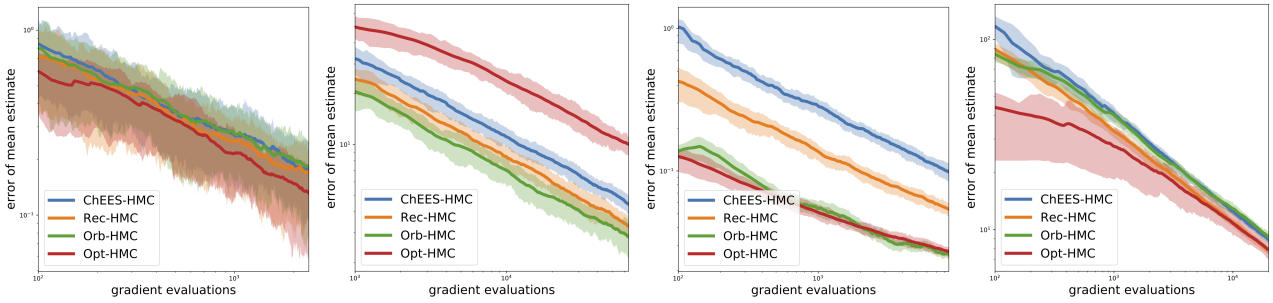

Figure 6: From left to right: the error of mean estimation on Banana, ill-conditioned Gaussian, logistic regression, Item-Response model. Every solid line depicts the mean of the absolute error averaged across 100 independent chains. The shaded area lies between 0.25 and 0.75 quantiles of the error.

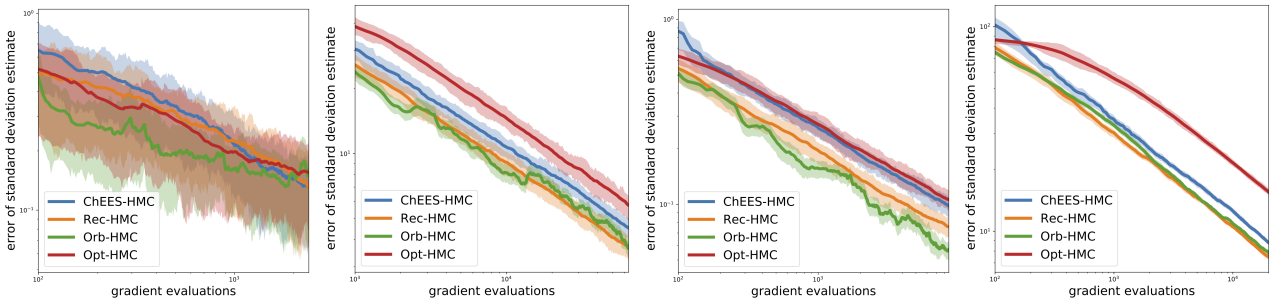

Figure 7: From left to right: the error of the variance estimation on Banana, ill-conditioned Gaussian, logistic regression, Item-Response model. Every solid line depicts the mean of the absolute error averaged across 100 independent chains. The shaded area lies between 0.25 and 0.75 quantiles of the error.

B.4 Distributions

Here we provide analytical forms of considered target distributions.

Banana (2-D):

$$p(x_1, x_2) = \mathcal{N}(x_1|0, 10)\mathcal{N}(x_2|0.03(x_1^2 - 100), 1) \tag{146}$$

Ill-conditioned Gaussian (50-D):

$$p(x) = \mathcal{N}(x|0, \Sigma), \tag{147}$$

where Σ is diagonal with variances from 10^{-2} to 10^2 in the log scale.

Bayesian logistic regression (25-D):
For the Bayesian logistic regression, we define likelihood and prior as

$$p(y = 1 \,|\, x, \theta) = \frac{1}{1 + \exp(-x^T\theta_w + \theta_b)}, \quad p(\theta) = \mathcal{N}(\theta \,|\, 0, \mathbb{1}). \tag{148}$$

Then the unnormalized density of the posterior distribution for a dataset $D = \{(x_i, y_i)\}_i$ is

$$p(\theta \,|\, D) \propto \prod_i p(y_i \,|\, x_i, \theta)p(\theta). \tag{149}$$

We use the German dataset (25 covariates, 1000 data points).

Item-response theory (501-D):
The model is defined by the joint distribution:

$$p(y, \alpha, \beta, \delta) = \prod_{n=1}^{N} \text{Bernoulli}(y_n \,|\, \sigma(\alpha_{j_n} - \beta_{k_n} + \delta)) \cdot \tag{150}$$

$$\cdot \prod_j \mathcal{N}(\alpha_j \,|\, 0, 1) \prod_k \mathcal{N}(\beta_k \,|\, 0, 1)\mathcal{N}(\delta \,|\, 0.75, 1). \tag{151}$$

Here α_j could be interpreted as the skill level of the student j; β_k is the difficulty of the question k; then y_n is the answer of a student to a question. We consider 100 students, 400 questions and 30105 responses. We generate the data from the prior.

B.5 Usage of neural models

In this section we discuss how one can possibly use the expressive learnable models (such as flows (Rezende & Mohamed, 2015; Dinh et al., 2016)) together with the orbital kernel to design an efficient sampler. As we discuss in Section 5 and Appendix B.3, to design an unbiased kernel one needs to design the periodic function. Therefore, in the next two subsection we consider the design of periodic functions using the family of normalizing flows.

B.5.1 Topologically conjugate sampler

This section operates similarly to (Hoffman et al., 2019). That is, having some simple function f and a learnable continuous bijection T one can sample using the topologically conjugate iterated function $h = T^{-1} \circ f \circ T$. Indeed, if the evaluation of f is cheap, then we can evaluate iterations of h as

$$h^n = T^{-1} \circ f^n \circ T. \tag{152}$$

For instance, we consider a simple periodic function f assuming that all the knowledge about the target distribution is encapsulated in T. Namely, we take Hamiltonian dynamics that can be integrated exactly:

$$H(x, v) = \frac{1}{2} x^T x + \frac{1}{2} v^T v, \tag{153}$$

$$x_i(\tau) = x_i(0) \cos(\tau) + v_i(0) \sin(\tau), \tag{154}$$

$$v_i(\tau) = -x_i(0) \sin(\tau) + v_i(0) \cos(\tau). \tag{155}$$

These equations define a continuous flow $f(x, \tau)$. To obtain discrete rotations, we can fix the evolution time either as $\tau = 2\pi \frac{1}{T}$ (obtaining periodic orbits with period T), or $\tau = 2\pi a$, where a is irrational number (obtaining returning orbits).

B.5.2 Periodic flows

Instead of the learning of the target space embedding, as in the previous section, we can learn the deterministic function f itself. For that purpose, we firstly introduce *periodic coupling layer*. Similarly to the coupling layer from (Dinh et al., 2016), we define the periodic coupling layer $l(x, \tau) : \mathbb{R}^D \to \mathbb{R}^D$ as follows.

$$y_{1:d} = x_{1:d} \tag{156}$$

$$y_{d+1:D} = x_{d+1:D} \odot \exp\Big(\sin(\omega\tau) s(x_{1:d})\Big) + \sin(\omega\tau) h(x_{1:d}) \tag{157}$$

Here $s(\cdot)$ and $h(\cdot)$ are neural networks that define scale and shift respectively; ω is the circular frequency, which is a hyperparameter. Stacking these layers we obtain the *periodic flow* $f(x, \tau)$, which is invertible for fixed τ and has tractable Jacobian. To guarantee periodicity, one can choose ω_i for each layer l_i as a natural number. Thus, we guarantee that $f(x, 0) = x$ and $f(x, 2\pi n) = x$ for natural n.

However, such flow does not satisfy $f(x, \tau_1 + \tau_2) = f(f(x, \tau_1), \tau_2)$. To ensure this property, we extend the state space by the auxiliary variable τ and introduce function $\widehat{f}((x, \tau), \Delta\tau) = [x', \tau']$ on the extended space:

$$\tau' = (\tau + \Delta\tau) \bmod 2\pi, \quad x' = f\Big(f^{-1}(x, \tau), \tau'\Big), \tag{158}$$

where the inversion of f is performed w.r.t. the x-argument. By the straightforward evaluation, we have

$$\widehat{f}((x, \tau), 0) = \widehat{f}((x, \tau), 2\pi) = (x, \tau), \quad \widehat{f}\Big(\widehat{f}((x, \tau), \Delta\tau_1), \Delta\tau_2\Big) = \widehat{f}((x, \tau), \Delta\tau_1 + \Delta\tau_2). \tag{159}$$

Finally, as in the previous section, we can obtain discrete rotations by considering $\Delta\tau = 2\pi \frac{1}{n}$ or $\Delta\tau = 2\pi a$ for irrational a.