Probabilistic Path Hamiltonian Monte Carlo

Vu Dinh [* 1]   **Arman Bilge** [* 1 2]   **Cheng Zhang** [* 1]   **Frederick A. Matsen IV** [1]

Abstract

Hamiltonian Monte Carlo (HMC) is an efficient and effective means of sampling posterior distributions on Euclidean space, which has been extended to manifolds with boundary. However, some applications require an extension to more general spaces. For example, phylogenetic (evolutionary) trees are defined in terms of both a discrete graph and associated continuous parameters; although one can represent these aspects using a single connected space, this rather complex space is not suitable for existing HMC algorithms. In this paper, we develop Probabilistic Path HMC (PPHMC) as a first step to sampling distributions on spaces with intricate combinatorial structure. We define PPHMC on orthant complexes, show that the resulting Markov chain is ergodic, and provide a promising implementation for the case of phylogenetic trees in open-source software. We also show that a surrogate function to ease the transition across a boundary on which the log-posterior has discontinuous derivatives can greatly improve efficiency.

1. Introduction

Hamiltonian Monte Carlo is a powerful sampling algorithm which has been shown to outperform many existing MCMC algorithms, especially in problems with high-dimensional and correlated distributions (Duane et al., 1987; Neal, 2011). The algorithm mimics the movement of a body balancing potential and kinetic energy by extending the state space to include auxiliary momentum variables and using Hamiltonian dynamics. By traversing long iso-probability contours in this extended state space, HMC is able to move long distances in state space in a single up-

date step, and thus has proved to be more effective than standard MCMC methods in a variety of applications. The method has gained a lot of interest from the scientific community and since then has been extended to tackle the problem of sampling on various geometric structures such as constrained spaces (Lan et al., 2014; Brubaker et al., 2012; Hartmann and Schütte, 2005), on general Hilbert space (Beskos et al., 2011) and on Riemannian manifolds (Girolami and Calderhead, 2011; Wang et al., 2013).

However, these extensions are not yet sufficient to apply to all sampling problems, such as in *phylogenetics*, the inference of evolutionary trees. Phylogenetics is the study of the evolutionary history and relationships among individuals or groups of organisms. In its statistical formulation it is an inference problem on hypotheses of shared history based on observed heritable traits under a model of evolution. Phylogenetics is an essential tool for understanding biological systems and is an important discipline of computational biology. The Bayesian paradigm is now commonly used to assess support for inferred tree structures or to test hypotheses that can be expressed in phylogenetic terms (Huelsenbeck et al., 2001).

Although the last several decades have seen an explosion of advanced methods for sampling from Bayesian posterior distributions, including HMC, phylogenetics still uses relatively classical Markov chain Monte Carlo (MCMC) based methods. This is in part because the number of possible tree topologies (the labeled graphs describing the branching structure of the evolutionary history) explodes combinatorially as the number of species increases. Also, to represent the phylogenetic relation among a fixed number of species, one needs to specify both the tree topology (a discrete object) and the branch lengths (continuous distances). This composite structure has thus far limited sampling methods to relatively classical Markov chain Monte Carlo (MCMC) based methods. One path forward is to use a construction of the set of phylogenetic trees as a single connected space composed of Euclidean spaces glued together in a combinatorial fashion (Kim, 2000; Moulton and Steel, 2004; Billera et al., 2001; Gavryushkin and Drummond, 2016) and try to define an HMC-like algorithm thereupon.

Experts in HMC are acutely aware of the need to extend HMC to such spaces with intricate combinatorial structure: Betancourt (2017) describes the extension of HMC to dis-

---

[*]Equal contribution [1]Program in Computational Biology, Fred Hutchison Cancer Research Center, Seattle, WA, USA [2]Department of Statistics, University of Washington, Seattle, WA, USA. Correspondence to: Frederick, A. Matsen IV <matsen@fredhutch.org>.
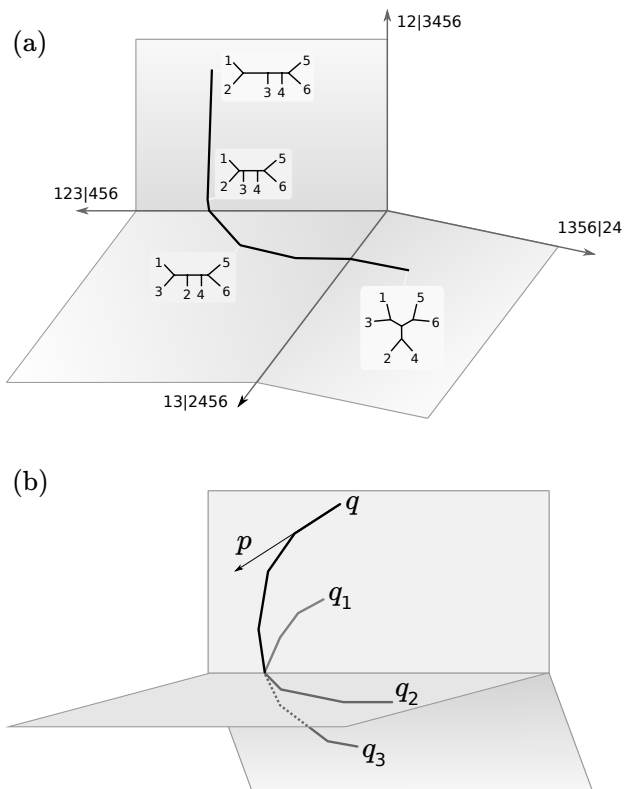
(a)

(b)

Figure 1. PPHMC on the orthant complex of tree space, in which each orthant (i.e. $\mathbb{R}^n_{\geq 0}$) represents the continuous branch length parameters of one tree topology. PPHMC uses the leapfrog algorithm to approximate Hamiltonian dynamics on each orthant, but can move between tree topologies by crossing boundaries between orthants. (a) A single PPHMC step moving through three topologies; each topology change along the path is one NNI move. (b) Because three orthants meet at every top-dimensional boundary, the algorithm must make a choice as to which topology to select. PPHMC uniformly selects a neighboring tree topology when the algorithm hits such a boundary. Here we show three potential outcomes q_1, q_2 and q_3 of running a single step of PPHMC started at position q with momentum p.

crete and tree spaces as a major outstanding challenge for the area. However, there are several challenges to defining a continuous dynamics-based sampling methods on such spaces. These tree spaces are composed of Euclidean components, one for each discrete tree topology, which are glued together in a way that respects natural similarities between trees. These similarities dictate that more than two such Euclidean spaces should get glued together along a common lower-dimensional boundary. The resulting lack of manifold structure poses a problem to the construction of an HMC sampling method on tree space, since up to now, HMC has just been defined on spaces with differential geometry. Similarly, while the posterior function is smooth within each topology, the function's behavior may be very

different between topologies. In fact, there is no general notion of differentiability of the posterior function on the whole tree space and any scheme to approximate Hamiltonian dynamics needs to take this issue into account.

In this paper, we develop Probabilistic Path Hamiltonian Monte Carlo (PPHMC) as a first step to sampling distributions on spaces with intricate combinatorial structure (Figure 1). After reviewing how the ensemble of phylogenetic trees is naturally described as a geometric structure we identify as an orthant complex (Billera et al., 2001), we define PPHMC for sampling posteriors on orthant complexes along with a probabilistic version of the leapfrog algorithm. This algorithm generalizes previous HMC algorithms by doing classical HMC on the Euclidean components of the orthant complex, but making random choices between alternative paths available at a boundary. We establish that the integrator retains the good theoretical properties of Hamiltonian dynamics in classical settings, namely probabilistic equivalents of time-reversibility, volume preservation, and accessibility, which combined together result in a proof of ergodicity for the resulting Markov chain. Although a direct application of the integrator to the phylogenetic posterior does work, we obtain significantly better performance by using a surrogate function near the boundary between topologies to control approximation error. This approach also addresses a general problem in using Reflective HMC (RHMC; Afshar and Domke, 2015) for energy functions with discontinuous derivatives (for which the accuracy of RHMC is of order $\mathcal{O}(\epsilon)$, instead of the standard local error $\mathcal{O}(\epsilon^3)$ of HMC on $\mathbb{R}^n$). We provide, validate, and benchmark two independent implementations in open-source software.

2. Mathematical framework

2.1. Bayesian learning on phylogenetic tree space

A *phylogenetic tree* (τ, q) is a tree graph τ with N leaves, each of which has a label, and such that each edge e is associated with a non-negative number q_e. Trees will be assumed to be bifurcating (internal nodes of degree 3) unless otherwise specified. We denote the number of edges of such a tree by $n = 2N - 3$. Any edge incident to a leaf is called a *pendant edge*, and any other edge is called an *internal edge*. Let $\mathcal{T}_N$ be the set of all N-leaf phylogenetic trees for which the lengths of pendant edges are bounded from below by some constant $e_0 > 0$. (This lower bound on branch lengths is a technical condition for theoretical development and can be relaxed in practice.)

We will use nearest neighbor interchange (NNI) moves (Robinson, 1971) to formalize what tree topologies that are "near" to each other. An NNI move is a transformation that collapses an internal edge to zero and then expands

the resulting degree 4 vertex into an edge and two degree 3 vertices in a new way (Figure 1a). Two tree topologies τ_1 and τ_2 are called *adjacent topologies* if there exists a single NNI move that transforms τ_1 into τ_2.

We will parameterize $\mathcal{T}_N$ as Billera-Holmes-Vogtmann (BHV) space (Billera et al., 2001), which we describe as follows. An *orthant* of dimension n is simply $\mathbb{R}^n_{\geq 0}$; each n-dimensional orthant is bounded by a collection of lower dimensional orthant *faces*. An *orthant complex* is a geometric object $\mathcal{X}$ obtained by gluing various orthants of the same dimension n, indexed by a countable set Γ, such that: (i) the intersection of any two orthants is a face of both orthants, and (ii) each $x \in \mathcal{X}$ belongs to a finite number of orthants. Each state of $\mathcal{X}$ is thus represented by a pair (τ, q), where $\tau \in \Gamma$ and $q \in \mathbb{R}^n_{\geq 0}$. Generalizing the definitions from phylogenetics, we refer to τ as its *topology* and to q as the vector of *attributes*. The topology of a point in an orthant complex indexes discrete structure, while the attributes formalize the continuous components of the space.

For phylogenetics, the complex is constructed by taking one n-dimensional orthant for each of the $(2n-3)!!$ possible tree topologies, and gluing them together along their common faces. The geometry can also be summarized as follows. In BHV space, each of these orthants parameterizes the set of branch lengths for a single topology (as a technical point, because we are bounding pendant branch lengths below by e_0, we can take the corresponding entries in the orthant to parameterize the amount of branch length above e_0). Top-dimensional orthants of the complex sharing a *facet*, i.e. a codimension 1 face, correspond to (NNI) adjacent topologies.

For a fixed phylogenetic tree (τ, q), the phylogenetic likelihood is defined as follows and will be denoted by $L(\tau, p)$ (see Kenney and Gu, 2012, for a full exposition). Let $\psi = (\psi_1, \psi_2, ..., \psi_S) \in \Omega^{N \times S}$ be the observed sequences (with characters in Ω) of length S over N leaves. The likelihood of observing ψ given the tree has the form

$$L(\tau, q) = \prod_{s=1}^{S} \sum_{a^s} \eta(a_\rho^s) \prod_{(u,v) \in E(\tau,q)} P_{a_u^s a_v^s}^{uv}(q_{uv})$$

where ρ is any internal node of the tree, a^s ranges over all extensions of ψ^s to the internal nodes of the tree, a_u^s denotes the assigned state of node u by a^s, $E(\tau, q)$ denotes the set of tree edges, $P_{ij}(t)$ denotes the transition probability from character i to character j across an edge of length t defined by a given evolutionary model and η is the stationary distribution of this evolutionary model. For this paper we will assume the simplest Jukes and Cantor (1969) model of a homogeneous continuous-time Markov chain on Ω with equal transition rates, noting that inferring parameters of complex substitution models is a vibrant yet separate subject of research (e.g. Zhao et al., 2016).

Given a proper prior distribution with density π_0 imposed on the branch lengths and on tree topologies, the posterior distribution can be computed as $\mathcal{P}(\tau, q) \propto L(\tau, q)\pi_0(\tau, q)$.

2.2. Bayesian learning on orthant complexes

With the motivation of phylogenetics in mind, we now describe how the phylogenetic problem sits as a specific case of a more general problem of Bayesian learning on orthant complexes, and distill the criteria needed to enable PPHMC on these spaces. This generality will also enable applications of Bayesian learning on similar spaces in other settings. For example in robotics, the state complex can be described by a cubical complexes whose vertices are the states, whose edges correspond to allowable moves, and whose cubes correspond to collections of moves which can be performed simultaneously (Ardila et al., 2014). Similarly, in learning on spaces of treelike shapes, the attributes are open curves translated to start at the origin, described by a fixed number of landmark points (Feragen et al., 2010).

An orthant complex, being a union of Euclidean orthants, naturally inherits the Lebesgue measure which we will denote hereafter by μ. Orthant complexes are typically not manifolds, thus to ensure consistency in movements across orthants, we assume that local coordinates of the orthants are defined in such a way that there is a natural one-to-one correspondence between the sets of attributes of any two orthants sharing a common face.

Assumption 2.1 (Consistency of local coordinates). *Given two topologies $\tau, \tau' \in \Gamma$ and state $x = (\tau, q_\tau) = (\tau', q_{\tau'})$ on the boundary of the orthants for τ and τ', then $q_\tau = q_{\tau'}$.*

We show that BHV tree space can be given such coordinates in the Appendix. For the rest of the paper, we define for each state $(\tau, q) \in \mathcal{X}$ the set $\mathcal{N}(\tau, q)$ of all *neighboring topologies* τ' such that τ' orthant contains (τ, q). Note that $\mathcal{N}(\tau, q)$ always includes τ, and if all coordinates of q are positive, $\mathcal{N}(\tau, q)$ is exactly $\{\tau\}$. Moreover, if $\tau' \in \mathcal{N}(\tau, q)$ and $\tau' \neq \tau$, we say that τ and τ' are *joined by* (τ, q). If the intersection of orthants for two topologies is a facet of each, we say that the two topologies are *adjacent*.

Finally, let $\mathcal{G}$ be the adjacency graph of the orthant complex $\mathcal{X}$, that is, the graph with vertices representing the topologies and edges connecting adjacent topologies. Recalling that the diameter of a graph is the maximum value of the graph distance between any two vertices, we assume that

Assumption 2.2. *The adjacency graph $\mathcal{G}$ of $\mathcal{X}$ has finite diameter, hereafter denoted by k.*

For phylogenetics, k is of order $\mathcal{O}(N \log N)$ (Li et al., 1996).

We seek to sample from a posterior distribution $\mathcal{P}(\tau, q)$ on $\mathcal{X}$. Assume that the negative logarithm of the posterior dis-

tribution $U(\tau, q) := -\log P(\tau, q)$ satisfies:

Assumption 2.3. $U(\tau, q)$ *is a continuous function on* $\mathcal{X}$, *and is smooth up to the boundary of each orthant* $\tau \in \Gamma$.

In the Appendix, we prove that if the logarithm of the phylogenetic prior distribution $\pi_0(\tau, q)$ satisfies Assumption 2.3, then so does the phylogenetic posterior distribution. It is also worth noting that while $U(\tau, q)$ is smooth within each orthant, the function's behavior may be very different between orthants and we do not assume any notion of differentiability of the posterior function on the whole space.

2.3. Hamiltonian dynamics on orthant complexes

The HMC state space includes auxiliary momentum variables in addition to the usual state variables. In our framework, the augmented state of this system is represented by the position (τ, q) and the momentum p, an n-dimensional vector. We will denote the set of all possible augmented state (τ, q, p) of the system by $\mathbb{T}$.

The Hamiltonian is defined as in the traditional HMC setting: $H(\tau, q, p) = U(\tau, q) + K(p)$, where $K(p) = \frac{1}{2}\|p\|^2$. We will refer to $U(\tau, q)$ and $K(p)$ as the potential energy function and the kinetic energy function of the system at the state (τ, q, p), respectively.

Our stochastic Hamiltonian-type system of equations is:

$$\frac{dp_i}{dt} = -\frac{\partial U}{\partial q_i}(\tau, q) \qquad \text{if } q_i > 0$$
$$p_i \leftarrow -p_i; \ \tau \sim Z(\mathcal{N}(\tau, q)) \qquad \text{if } q_i = 0 \qquad (2.1)$$
$$\frac{dq_i}{dt} = p_i$$

where $Z(A)$ denotes the uniform distribution on the set A.

If all coordinates of q are positive, the system behaves as in the traditional Hamiltonian setting on $\mathbb{R}^n$. When some attributes hit zero, however, the new orthant is picked randomly from the orthants of neighboring topologies (including the current one), and the momenta corresponding to non-positive attributes are negated.

Assumption 2.3 implies that despite the non-differentiable changes in the governing equation across orthants, the Hamiltonian of the system along any path is constant:

Lemma 2.1. H *is conserved along any system dynamics.*

2.4. A probabilistic "leap-prog" algorithm

In practice, we approximate Hamiltonian dynamics by the following integrator with step size ϵ, which we call "leap-prog" as a probabilistic analog of the usual leapfrog algorithm. This extends previous work of (Afshar and Domke, 2015) on RHMC where particles can reflect against planar surfaces of $\mathbb{R}^n_{\geq 0}$.

In the RHMC formulation, one breaks the step size ϵ into smaller sub-steps, each of which correspond to an event when some of the coordinates cross zero. We adapt this idea to HMC on orthant complexes as follows. Every time such an event happens, we reevaluate the values of the position and the momentum vectors, update the topology (uniformly at random from the set of neighboring topologies), reverse the momentum of crossing coordinates and continue the process until a total step size ϵ is achieved (Algorithm 1). We note that several topologies might be visited in one leap-prog step.

If there are no topological changes in the trajectory to time ϵ, this procedure is equivalent to classical HMC. Moreover, since the algorithm only re-evaluates the gradient of the energy function at the end of the step when the final position has been fixed, changes in topology on the path have no effect on the changes of position and momentum. Thus, the projection of the particles (in a single leap-prog step) to the (q, p) space is identical to a leapfrog step of RHMC on $\mathbb{R}^n_{\geq 0}$.

Algorithm 1 Leap-prog algorithm with step size ϵ.

> $p \leftarrow p - \epsilon \nabla U(\tau, q)/2$
> **if** FirstUpdateEvent$(\tau, q, p, \epsilon) = \emptyset$ **then**
> $\quad q \leftarrow q + \epsilon p$
> **else**
> $\quad t \leftarrow 0$
> $\quad$ **while** FirstUpdateEvent$(\tau, q, p, \epsilon - t) \neq \emptyset$ **do**
> $\quad\quad (q, e, I) \leftarrow$ FirstUpdateEvent$(\tau, q, p, \epsilon - t)$
> $\quad\quad t \leftarrow t + e$
> $\quad\quad \tau \sim Z(\mathcal{N}(\tau, q))$
> $\quad\quad p_I \leftarrow -p_I$
> $\quad$ **end while**
> $\quad q \leftarrow q + (\epsilon - t)p$
> **end if**
> $p \leftarrow p - \epsilon \nabla U(\tau, q)/2$

Here FirstUpdateEvent(τ, q, p, t) returns x, the position of the first event for which the line segment $[q, q + tp]$ crosses zero; e, the time when this event happens; and I, the indices of the coordinates crossing zero during this event. If q_i and p_i are both zero before FirstUpdateEvent is called, i is not considered as a crossing coordinate. If no such an event exists, $\emptyset$ is returned.

3. Hamiltonian Monte Carlo on orthant complexes

Probabilistic Path Hamiltonian Monte Carlo (PPHMC) with leap-prog dynamics iterates three steps, similar to that of classical HMC. First, new values for the momentum variables are randomly drawn from their Gaussian distribution, independently of the current values of the posi-

tion variables. Second, starting with the current augmented state, $s = (\tau, q, p)$, the Hamiltonian dynamics is run for a fixed number of steps T using the leap-prog algorithm with fixed step size ϵ. The momentum at the end of this trajectory is then negated, giving a proposed augmented state $s^* = (\tau^*, q^*, p^*)$. Finally, this proposed augmented state is accepted as the next augmented state of the Markov chain with probability $r(s, s^*) = \min\left(1, \exp(H(s) - H(s^*))\right)$.

PPHMC has two natural advantages over MCMC methods for phylogenetic inference: jumps between topologies are guided by the potential surface, and many jumps can be combined into a single proposal with high acceptance probability. Indeed, rather than completely random jumps as used for MCMC, the topological modifications of HMC are guided by the gradient of the potential. This is important because there are an enormous number of phylogenetic trees, namely $(2n - 3)!!$ trees with n leaves. Secondly, HMC can combine a great number of tree modifications into a single step, allowing for large jumps in tree space with high acceptance probability. These two characteristics are analogs of why HMC is superior to conventional MCMC in continuous settings and what we aimed to extend to our problem.

3.1. Theoretical properties of the leap-prog integrator

To support the use of this leap-prog integrator for MCMC sampling, we establish that integrator retains analogs of the good theoretical properties of Hamiltonian dynamics in classical settings, namely, time-reversibility, volume preservation and accessibility (proofs in Appendix).

We formulate probabilistic reversibility as:

Lemma 3.1 (Reversibility). *For a fixed finite time horizon T, we denote by $P(s, s')$ the probability that the integrator moves s to s' in a single update step. We have*

$$P((\tau, q, p), (\tau', q', p')) = P((\tau', q', -p'), (\tau, q, -p)).$$

for any augmented states (τ', q', p') and $(\tau, q, p) \in \mathbb{T}$.

The central part of proving the detailed balance condition of PPHMC is to verify that Hamiltonian dynamics preserves volume. Unlike the traditional HMC setting where the proposal distribution is a single point mass, in our probabilistic setting, if we start at one augmented state s, we may end up at countably many end points due to stochastic HMC integration. The equation describing volume preservation in this case needs to be generalized to the form of Equation (3.1), where the summations account for the discreteness of the proposal distribution.

Lemma 3.2 (Volume preservation). *For every pair of measurable sets $A, B \subset \mathbb{T}$ and elements $s, s' \in \mathbb{T}$, we denote by $P(s, s')$ the probability that the integrator moves s to s'*

in a single update step and define

$$B(s) = \{s' \in B : P(s, s') > 0\}$$

and

$$A(s') = \{s \in A : P(s, s') > 0\}.$$

Then

$$\int_A \sum_{s' \in B(s)} P(s, s')\, ds = \int_B \sum_{s \in A(s')} P(s', s)\, ds'. \quad (3.1)$$

If we restrict to the case of trajectories staying in a single topology, $A(s)$ and $B(s')$ are singletons and we get back the traditional equation of volume preservation. We also note that the measure ds in Equation (3.1) is the Lebesgue measure: when there is no randomness in the Hamiltonian paths, (3.1) becomes the standard volume preservation condition, where volumes are expressed by the Lebesgue measure.

Typically, accessibility poses no major problem in various settings of HMC since it is usually clear that one can go between any two positions in a single HMC step. In the case of PPHMC, however, the composition of discrete and continuous structure, along with the possible non-differentiability of the potential energy across orthants, make it challenging to verify this condition. Here we show instead that the PPHMC algorithm can go between any two states with k steps, where k denotes the diameter of adjacency graph $\mathcal{G}$ the space $\mathcal{X}$ and each PPHMC step consists of T leap-prog steps of size ϵ.

Lemma 3.3 (k-accessibility). *For a fixed starting state $(\tau^{(0)}, q^{(0)})$, any state $(\tau', q') \in \mathcal{X}$ can be reached from $(\tau^{(0)}, q^{(0)})$ by running k steps of PPHMC.*

The proof of this Lemma is based on Assumption 2.2, which asserts that the adjacency graph $\mathcal{G}$ of $\mathcal{X}$ has finite diameter, and that classical HMC allows the particles to move freely in each orthant by a single HMC step.

To show that Markov chains generated by PPHMC are ergodic, we also need to prove that the integrator can reach any subset of positive measure of the augmented state space with positive probability. To enable such a result, we show:

Lemma 3.4. *For every sequence of topologies $\omega = \{\tau^{(0)}, \tau^{(1)}, \ldots, \tau^{(n_\omega)}\}$ and every set with positive measure $B \subset \mathcal{X}$, let B_ω be the set of all $(\tau', q') \in B$ such that (τ', q') can be reached from $(\tau^{(0)}, q^{(0)})$ in k PPHMC steps and such that the sequence of topologies crossed by the trajectory is ω. We denote by $I_{B,\omega}$ the set of all sequences of initial momenta for each PPHMC step $\{p^{(0)}, \ldots, p^{(k)}\}$ that make such a path possible.*

Then, if $\mu(I_{B,\omega}) = 0$, then $\mu(B_\omega) = 0$.

We also need certain sets to be countable.

Lemma 3.5. *Given $s \in \mathbb{T}$, we denote by $R(s)$ the set of all augmented states s' such that there is a finite-size leap-prog step with path γ connecting s and s', and by $K(s)$ the set of all such leap-prog paths γ connecting s and $s' \in R(s)$. Then $R(s)$ and $K(s)$ are countable. Moreover, the probability $P_\infty(s, s')$ of moving from s to s' via paths with infinite number of topological changes is zero.*

3.2. Ergodicity of Probabilistic Path HMC

In this section, we establish that a Markov chain generated by PPHMC is ergodic with stationary distribution $\pi(\tau, q) \propto \exp(-U(\tau, q))$. To do so, we need to verify that the Markov chain generated by PPHMC is aperiodic, because we have shown k-accessibility of the integrator rather than 1-accessibility. Throughout this section, we will use the notation $P((\tau, q, p), \cdot)$ to denote the one-step proposal distribution of PPHMC starting at augmented state (τ, q, p), and $P((\tau, q), \cdot)$ to denote the one-step proposal distribution of PPHMC starting at position (τ, q) and with a momentum vector drawn from a Gaussian as described above.

We first note that:

Lemma 3.6. *PPHMC preserves the target distribution π.*

Given probabilistic volume preservation (3.2), the proof is standard and is given in the Appendix.

Theorem 3.1 (Ergodic)**.** *The Markov chain generated by PPHMC is ergodic.*

Proof of Theorem 3.1. For every sequence of topologies $\omega = \{\tau^{(0)}, \tau^{(1)}, \ldots, \tau^{(n_\omega)}\}$ (finite by Lemma 3.5) and every set with positive measure $B \subset \mathcal{X}$, we define B_ω and $I_{B,\omega}$ as in the previous section. By Lemma 3.3, we have

$$B = \bigcup_\omega B_\omega.$$

Assume that $\mu(I_{B,\omega}) = 0$ for all ω. From Lemma 3.4, we deduce that $\mu(B_\omega) = 0$ for all ω. This makes $\mu(B) = 0$, which is a contradiction. Hence $\mu(I_{B,\omega}) > 0$ for some ω and $P^{n_\omega}((\tau^{(0)}, q^{(0)}), B)$ is at least the positive quantity

$$\frac{1}{Z} \int_{p \in I_{B,\omega}} P^{n_\omega}((\tau^{(0)}, q^{(0)}, p), B) \, \exp(-K(p)) dp$$

where Z is the normalizing constant. This holds for all sets with positive measure $B \subset \mathcal{X}$, so PPHMC is irreducible.

Now assume that a Markov chain generated by the leapfrog algorithm is periodic. The reversibility of Hamiltonian dynamics implies that the period d must be equal to 2. In other words, there exist two disjoint subsets X_1, X_2 of $\mathcal{X}$ such that $\pi(X_1) > 0$, and

$$P(x, X_2) = 1 \;\; \forall x \in X_1, \text{ and } \;\; P(x, X_1) = 1 \;\; \forall x \in X_2.$$

Consider $x \in X_1$ with all positive attributes. There exists a neighborhood U_x around x such that any $y \in U_x$ is reachable from x by Hamiltonian dynamics. Since X_1, X_2 are disjoint, we deduce that $\mu(U_x \cap X_1) = 0$. Since the neighborhood U_x exists for almost every $x \in X_1$, this implies that $\mu(X_1) = 0$, and hence, that $\pi(X_1) = 0$, which is a contradiction. We conclude that any Markov chain generated by the leapfrog algorithm is aperiodic.

Lemma 3.6 shows that PPHMC preserves the target distribution π. This, along with π-irreducibility and aperiodicity, completes the proof (Roberts and Rosenthal, 2004). $\square$

3.3. An efficient surrogate smoothing strategy

One major advantage of HMC methods over traditional approaches is that HMC-proposed states may be distant from the current state but nevertheless have a high probability of acceptance. This partially relies on the fact that the leapfrog algorithm with smooth energy functions has a local approximation error of order $\mathcal{O}(\epsilon^3)$ (which leads to global error $\mathcal{O}(T\epsilon^3)$, where T is the number of leapfrog steps in a Hamiltonian path).

However, when the potential energy function $U(\tau, q)$ is not differentiable on the whole space this low approximation error can break down. Indeed, although PPHMC inherits many nice properties from vanilla HMC and RHMC (Afshar and Domke, 2015), this discontinuity of the derivatives of the potential energy across orthants may result in non-negligible loss of accuracy during numerical simulations of the Hamiltonian dynamics. A careful analysis of the local approximation error of RHMC for potential energy functions with discontinuous first derivatives reveals that it only has an local error rate of order at least $\Omega(\epsilon)$ (see proof in Appendix):

Proposition 3.1. *Given a potential function V, we denote by V^+ and V^- the restrictions of V on the half-spaces $\{x_1 \geq 0\}$ and $\{x_1 \leq 0\}$ and assume that V^+ and V^- are smooth up to the boundary of their domains. If the first derivative with respect to the first component of the potential energy $V(q)$ are discontinuous across the hyperplane $\{x_1 = 0\}$ (i.e., $(\partial V^+)/(\partial q_1)$ and $(\partial V^-)/(\partial q_1)$ are not identical on this set), then RHMC on this hyper-plane has a local error of order at least $\Omega(\epsilon)$.*

Since PPHMC uses RHMC, when the first derivatives of the potential energy are discontinuous, it also has a global error of order $\mathcal{O}(C\epsilon + T\epsilon^3)$, which depends on the number of *critical events* C along a Hamiltonian path (that is, the number of reflection/refraction events). This makes it difficult to tune the step size ϵ for optimal acceptance rate, and requires small ϵ, limiting topology exploration.

To alleviate this issue, we propose the use of a surrogate induced Hamiltonian dynamics (Strathmann et al., 2015;

Zhang et al., 2016) with the Hamiltonian $\tilde{H}(\tau, q, p) = \tilde{U}(\tau, q) + K(p)$, where the surrogate potential energy is

$$\tilde{U}(\tau, q) = U(\tau, G(q)), \quad G(q) = (g(q_1), \ldots, g(q_n))$$

and $g(x)$ is some positive and smooth approximation of $|x|$ with vanishing gradient at $x = 0$. One simple example which will be used for the rest of this paper is

$$g_\delta(x) = \begin{cases} x, & x \geq \delta \\ \frac{1}{2\delta}(x^2 + \delta^2), & 0 \leq x < \delta \end{cases}$$

where δ will be called the *smoothing threshold*.

Due to the vanishing gradient of g, $\tilde{U}$ now has continuous derivatives across orthants. However, $\tilde{U}$ is no longer continuous across orthants since $g(0) \neq 0$ and we thus employ the refraction technique introduced in Afshar and Domke (2015) (see Algorithm 2 for more details). The proposed state $s_\delta^* = (\tau_\delta^*, q_\delta^*, p_\delta^*)$ at the end of the trajectory is accepted with probability according to the original Hamiltonian, that is, $\min(1, \exp(H(s) - H(s_\delta^*)))$.

By following the same framework proposed in previous sections, we can prove that the resulting sampler still samples from the exact posterior distribution $\mathcal{P}(\tau, q)$. A complete treatment, however, requires more technical adjustments and is beyond the scope of the paper. We will leave this as a subject of future work.

Algorithm 2 Refractive Leap-prog with surrogate

$p \leftarrow p - \epsilon \nabla \tilde{U}(\tau, q)/2$
if FirstUpdateEvent$(\tau, q, p, \epsilon) = \emptyset$ **then**
    $q \leftarrow q + \epsilon p$
else
    $t \leftarrow 0$
    while FirstUpdateEvent$(q, p, \epsilon - t) \neq \emptyset$ **do**
        $(q, e, I) \leftarrow$ FirstUpdateEvent$(\tau, q, p, \epsilon - t)$
        $t \leftarrow t + e$
        $\tau' \sim Z(\mathcal{N}(\tau, q))$
        $\Delta E \leftarrow \tilde{U}(\tau', q) - \tilde{U}(\tau, q)$
        if $\|p_I\|^2 > 2\Delta E$ **then**
            $p_I \leftarrow \sqrt{\|p_I\|^2 - 2\Delta E} \cdot \dfrac{-p_I}{\|p_I\|}$
            $\tau \leftarrow \tau'$
        else
            $p_I \leftarrow -p_I$
        end if
    end while
    $q \leftarrow q + (\epsilon - t)p$
end if
$p \leftarrow p - \epsilon \nabla \tilde{U}(\tau, q)/2$

As we will illustrate later, compared to the exact potential energy, the continuity of the derivative of the surrogate potential across orthants dramatically reduces the discretiza-

tion error and allows for high acceptance probability with relatively large step size.

4. Experiments

In this section, we demonstrate the validity and efficiency of our PPHMC method by an application to Bayesian phylogenetic inference. We compared our PPHMC implementations to industry-standard MrBayes 3.2.5, which uses MCMC to sample phylogenetic trees (Ronquist et al., 2012). We concentrated on the most challenging part: sampling jointly for the branch lengths and tree topologies, and assumed other parameters (e.g., substitution model, hyper-parameters for the priors) are fixed. More specifically, for all of our experiments we continued to assume the Jukes-Cantor model of DNA substitution and placed a uniform prior on the tree topology $\tau \sim Z(\mathcal{T}_N)$ with branch lengths i.i.d. $q_i \sim \text{Exponential}(\lambda = 10)$, as done by others when measuring the performance of MCMC algorithms for Bayesian phylogenetics (e.g., Whidden and Matsen, 2015). As mentioned earlier, although in the theoretical development we assumed that the lengths of the pendant edges are bounded from below by a positive constant e_0 to ensure that the likelihood stays positive on the whole tree space, this condition is not necessary in practice since the Hamiltonian dynamics guide the particles away from regions with zero likelihood (i.e., the region with $U = \infty$). We validate the algorithm through two independent implementations in open-source software:

1. a Scala version available at https://github.com/armanbilge/phyloHMC that uses the Phylogenetic Likelihood Library[1] (Flouri et al., 2015), and

2. a Python version available at https://github.com/zcrabbit/PhyloInfer that uses the ETE toolkit (Huerta-Cepas et al., 2016) and Biopython (Cock et al., 2009).

4.1. Simulated data

As a proof of concept, we first tested our PPHMC method on a simulated data set. We used a random unrooted tree with $N = 50$ leaves sampled from the aforementioned prior. 1000 nucleotide observations for each leaf were then generated by simulating the continuous-time Markov model along the tree. This moderate data set provided enough information for model inference while allowing for a relatively rich posterior distribution to sample from.

We ran MrBayes for 10^7 iterations and sampled every 1000 iterations after a burn-in period of the first 25% iterations to establish a ground truth for the posterior distribution.

---

[1] https://github.com/xflouris/libpll

For PPHMC, we set the step size $\epsilon = 0.0015$ and smoothing threshold $\delta = 0.003$ to give an overall acceptance rate of about $\alpha = 0.68$ and set the number of leap-prog steps $T = 200$. We then ran PPHMC for $10,000$ iterations with a burn-in of 25%. We saw that PPHMC indeed samples from the correct posterior distribution (see Figure S1 in the Appendix).

4.2. Empirical data

We also analyzed an empirical data set labeled DS4 by Whidden and Matsen (2015) that has become a standard benchmark for MCMC algorithms for Bayesian phylogenetics since Lakner et al. (2008). DS4 consists of 1137 nucleotide observations per leaf from $N = 41$ leaves representing different species of fungi. Notably, only 554 of these observations are complete; the remaining 583 are missing a character for one or more leaves so the likelihood is marginalized over all possible characters. Whidden and Matsen (2015) observed that the posterior distribution for DS4 features high-probability trees separated by paths through low-probability trees and thus denoted it a "peaky" data set that was difficult to sample from using MrBayes.

To find the optimal choice of tuning parameters for DS4, we did a grid search on the space of step size ϵ and the smoothing threshold–step size ratio δ/ϵ. The number of leap-prog steps T was adjusted to keep the total simulation time ϵT fixed. For each choice of parameters, we estimated the expected acceptance rate α by averaging over 20 proposals from the PPHMC transition kernel per state for 100 states sampled from the posterior in a previous, well-mixed run. This strategy enabled us to obtain an accurate estimate of the average acceptance rate without needing to account for the different mixing rates in a full PPHMC run due to the various settings for the tuning parameters.

The results suggest that choosing $\delta \approx 2\epsilon$ maximizes the acceptance probability (Figure 2b). Furthermore, when aiming for the optimal acceptance rate of $\alpha = 0.65$ (Neal, 2011), the use of the surrogate function enables a choice of step size ϵ nearly 10 times greater than otherwise. In practice, this means that an equivalent proposal requires less leap-prog steps and gives a more efficient sampling algorithm.

To see the difference this makes in practice, we ran long trajectories for exact and surrogate-smoothed PPHMC with a relatively large step size $\epsilon = 0.0008$. Indeed, we found that the surrogate enables very long trajectories and large number of topology transformations (Figure 2a,c).

5. Conclusion

Sophisticated techniques for sampling posteriors using HMC have thus far been restricted to manifolds with
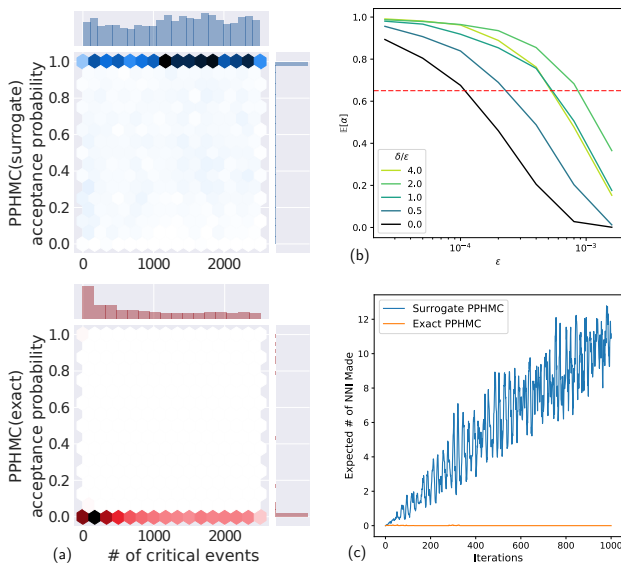

Figure 2. (a) A hexbin plot comparison showing that the surrogate-smoothed PPHMC achieves much higher acceptance rate and longer paths than exact PPHMC on the DS4 data set. (b) Average acceptance rate α for different choices of step size ϵ and smoothing threshold δ for phylogenetic HMC applied to DS4, where $\epsilon T = 0.08$. The black series ($\delta/\epsilon = 0.0$) is equivalent to exact PPHMC. The dashed red line indicates the acceptance rate $\alpha = 0.65$. (c) Expected number of NNI moves made by both exact and surrogate-smoothed PPHMC. Results are obtained by averaging 100 HMC iterations with long trajectories ($T = 1000$), starting from the same posterior sample.

boundary. To address this limitation, we have developed "PPHMC," which is the first extension of HMC to a space with intricate combinatorial structure. The corresponding integrator makes a random choice among alternatives when encountering a boundary. To prove ergodicity, we extend familiar elements of HMC proofs to this probabilistic path setting. We develop a smoothing surrogate function that enables long HMC paths with many boundary transitions across which the posterior is not differentiable. Our surrogate method enables high acceptance probability for RHMC (Afshar and Domke, 2015) in the case of potential functions with discontinuous derivatives; this aspect of our work is independent of the probabilistic nature of PPHMC. Our implementation shows good performance on both simulated and real data. There are many opportunities for future development, including extending the theory to other classes of combinatorially-described spaces and surrogate functions, developing adaptive path length algorithms, as well as extending our implementation for phylogenetics to sample other mutation model and demographic parameters along with topologies.

References

Hadi Mohasel Afshar and Justin Domke. Reflection, refraction, and Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 2989–2997, 2015.

Federico Ardila, Tia Baker, and Rika Yatchak. Moving robots efficiently using the combinatorics of cat (0) cubical complexes. *SIAM Journal on Discrete Mathematics*, 28(2):986–1007, 2014.

Alexandros Beskos, Frank J Pinski, Jesús María Sanz-Serna, and Andrew M Stuart. Hybrid Monte Carlo on Hilbert spaces. *Stochastic Processes and their Applications*, 121(10):2201–2230, 2011.

Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Louis J Billera, Susan P Holmes, and Karen Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767, 2001.

Marcus A Brubaker, Mathieu Salzmann, and Raquel Urtasun. A family of MCMC methods on implicitly defined manifolds. In *International Conference on Artificial Intelligence and Statistics*, pages 161–172, 2012.

David Bryant. The splits in the neighborhood of a tree. *Ann. Comb.*, 8(1):1–11, 2004. ISSN 0218-0006. URL http://link.springer.com/article/10.1007/s00026-004-0200-z.

O Peter Buneman. The recovery of trees from measures of dissimilarity. *Mathematics in the archaeological and historical sciences*, 1971. URL http://ci.nii.ac.jp/naid/10018379872/.

Eric Cances, Frédéric Legoll, and Gabriel Stoltz. Theoretical and numerical comparison of some sampling methods for molecular dynamics. *ESAIM: Mathematical Modelling and Numerical Analysis*, 41(02):351–389, 2007.

P. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25:1422–1423, 2009.

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

Lawrence Craig Evans and Ronald F Gariepy. *Measure theory and fine properties of functions*. CRC press, 2015.

Aasa Feragen, Francois Lauze, Pechin Lo, Marleen de Bruijne, and Mads Nielsen. Geometries on spaces of treelike shapes. In *Asian Conference on Computer Vision*, pages 160–173. Springer, 2010.

T. Flouri, F. Izquierdo-Carrasco, D. Darriba, A.J. Aberer, L.-T. Nguyen, B.Q. Minh, A. Von Haeseler, and A. Stamatakis. The phylogenetic likelihood library. *Syst. Bio.*, 64(2):356–362, 2015. doi: 10.1093/sysbio/syu084.

Alex Gavryushkin and Alexei J Drummond. The space of ultrametric phylogenetic trees. *J. Theor. Biol.*, 403:197–208, 21 August 2016. ISSN 0022-5193, 1095-8541. doi: 10.1016/j.jtbi.2016.05.001. URL http://dx.doi.org/10.1016/j.jtbi.2016.05.001.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Carsten Hartmann and Christof Schütte. A constrained hybrid Monte Carlo algorithm and the problem of calculating the free energy in several variables. *ZAMM-Journal of Applied Mathematics and Mechanics*, 85(10):700–710, 2005.

J P Huelsenbeck, F Ronquist, R Nielsen, and J P Bollback. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294(5550):2310–2314, 14 December 2001. ISSN 0036-8075. doi: 10.1126/science.1065889. URL http://dx.doi.org/10.1126/science.1065889.

Jaime Huerta-Cepas, Francois Serra, and Peek Bork. ETE3: Reconstruction, analysis and visualization of phylogenomic data. *Mol. Biol. Evol.*, 2016.

Thomas H. Jukes and Charles R. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mammalian protein metabolism*, pages 21–132. Academic Press, New York, 1969.

Toby Kenney and Hong Gu. Hessian calculation for phylogenetic likelihood based on the pruning algorithm and its applications. *Stat. Appl. Genet. Mol. Biol.*, 11(4):Article 14, 25 September 2012. ISSN 1544-6115. doi: 10.1515/1544-6115.1779. URL http://dx.doi.org/10.1515/1544-6115.1779.

Junhyong Kim. Slicing hyperdimensional oranges: the geometry of phylogenetic estimation. *Molecular phylogenetics and evolution*, 17(1):58–75, 2000.

Clemens Lakner, Paul van der Mark, John P Huelsenbeck, Bret Larget, and Fredrik Ronquist. Efficiency of markov chain monte carlo tree proposals in bayesian phylogenetics. *Syst. Biol.*, 57(1):86–103, February 2008. ISSN 1063-5157. doi: 10.1080/10635150801886156. URL http://dx.doi.org/10.1080/10635150801886156.

Shiwei Lan, Bo Zhou, and Babak Shahbaba. Spherical Hamiltonian Monte Carlo for constrained target distributions. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 629–637, 2014.

M Li, J Tromp, and L Zhang. On the nearest neighbour interchange distance between evolutionary trees. *J. Theor. Biol.*, 182(4):463–467, 21 October 1996. ISSN 0022-5193. doi: 10.1006/jtbi.1996.0188. URL http://dx.doi.org/10.1006/jtbi.1996.0188.

Vincent Moulton and Mike Steel. Peeling phylogenetic oranges. *Advances in Applied Mathematics*, 33(4):710–727, 2004.

Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2, 2011.

Gareth O Roberts and Jeffrey S Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

David F Robinson. Comparison of labeled trees with valency three. *Journal of Combinatorial Theory, Series B*, 11(2):105–119, 1971.

Fredrik Ronquist, Maxim Teslenko, Paul van der Mark, Daniel L Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A Suchard, and John P Huelsenbeck. MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.*, 61(3):539–542, 22 February 2012. ISSN 1063-5157. doi: 10.1093/ sysbio/sys029. URL http://dx.doi.org/10.1093/ sysbio/sys029.

C Semple and M Steel. *Phylogenetics*. Oxford University Press, New York, NY, 2003.

Heiko Strathmann, Dino Sejdinovic, Samuel Livingstone, Zoltan Szabo, and Arthur Gretton. Gradient-free Hamiltonian Monte Carlo with efficient kernel exponential families. In *Advances in Neural Information Processing Systems*, pages 955–963, 2015.

Ziyu Wang, Shakir Mohamed, and Nando de Freitas. Adaptive Hamiltonian and Riemann manifold monte carlo. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1462–1470, 2013.

Chris Whidden and IV Matsen, Frederick A. Quantifying MCMC exploration of phylogenetic tree space. *Syst. Biol.*, 64(3):472–491, 2015.

Cheng Zhang, Babak Shahbaba, and Hongkai Zhao. Hamiltonian Monte Carlo acceleration using surrogate functions with random bases. *Statistics and Computing*, pages 1–18, 2016.

Tingting Zhao, Ziyu Wang, Alexander Cumberworth, Joerg Gsponer, Nando de Freitas, and Alexandre Bouchard-Côté. Bayesian analysis of continuous time markov chains with application to phylogenetic modelling. *Bayesian Anal.*, 2016. ISSN 1936-0975, 1931-6690. URL http:// projecteuclid.org/euclid.ba/1448899900.