CONVERGENCE OF RANDOM WALKS TO BROWNIAN MOTION IN PHYLOGENETIC TREE-SPACE

TOM M. W. NYE

ABSTRACT. The set of all phylogenetic trees for a fixed set of species forms a geodesic metric space known as Billera-Holmes-Vogtmann tree-space. In order to analyse samples of phylogenetic trees it is desirable to construct parametric distributions on this space, but this task is very challenging. One way to construct such distributions is to consider particles undergoing Brownian motion in tree-space from a fixed starting point. The distribution of the particles after a given duration of time is analogous to a multivariate normal distribution in Euclidean space. Since these distributions cannot be worked with directly, we consider approximating by suitably defined random walks in tree-space. We prove that as the number of steps tends to infinity and the step-size tends to zero, the distributions obtained by random walk converge to those corresponding to Brownian motion. This result opens the possibility of statistical modelling using distributions obtained from Brownian motion on tree-space.

1. Introduction

Phylogenetic trees represent evolutionary relationships between a set $S$ of biological objects, typically genetic sequences obtained from different present-day species. Internal vertices correspond to historic divergence events and the leaf vertices correspond to elements in $S$. The tree are edge-weighted, and these weights represent the degree of evolutionary divergence between species. The set of all phylogenetic trees for a fixed set $S$ forms a geodesic metric space usually referred to as Billera-Holmes-Vogtmann (BHV) tree-space \cite{2}. Samples of trees arise from many biological analyses, for example by constructing a phylogenetic tree for each gene in a collection of different genes present in all species $S$. Medical images of branching structures in the body, such as blood vessels \cite{10} and lung airways \cite{4}, can also be represented as phylogenetic trees, and samples of images have been analysed using methods based in BHV tree-space.

To date, most methods for analysing data sets in BHV tree-space have involved least squares estimation using the geodesic metric: there are methods for computing sample means \cite{1, 7} and for fitting principal geodesics \cite{4, 8}, for example. On the other hand, development of fully probabilistic models is difficult due to the challenging problem of construction of useful probability distributions on tree-space. Some recent work has focussed on kernel density estimation for non-parametric estimation of distributions on tree-space \cite{11}. In this article we consider stochastic processes on tree-space, and use Brownian motion to construct probability distributions. Consider particles undergoing Brownian motion from some fixed point $x_0$ in tree-space: if we run the Brownian motion for some fixed time duration $t_0$ then the corresponding distribution of particles will be denoted $B(x_0, t_0)$. This distribution consists of a ‘bump’ of density on tree-space characterized by the location
of the source $x_0$ and the time $t_0$ which plays the role of a dispersion parameter. If the same process runs in Euclidean space then the corresponding distribution is the multivariate normal distribution with mean $x_0$ and variance-covariance matrix $t_0 \times I$. The density function for $B(x_0, t_0)$ for the space of rooted phylogenies has been explicitly calculated in the cases of $|S| = 3$ and 4 species [9]. However, for $|S| > 4$ calculation of the density function becomes very difficult since it involves gluing together spherical harmonics and higher dimensional analogues to create eigenfunctions of the Laplacian on tree-space. Rather than attempting to calculate the density function explicitly, an alternative is to approximate Brownian motion using a suitably defined random walk. This has the additional benefit that random walks are easy to simulate, and it opens the possibility of using stochastic processes on tree-space more generally.

In this paper we address two fundamental issues: (i) existence of Brownian motion as a well-defined Markov process on tree-space, and (ii) convergence of random walks to Brownian motion under a certain limit. BHV tree-space is a manifold stratified space: it consists of a collection of regions which are isomorphic to the positive orthant in $\mathbb{R}^n$, with one such region for each labelled tree shape. These regions are glued along their faces in a way determined by the combinatorics of the tree shapes. Brownian motion on tree-space proceeds in the same way as for $\mathbb{R}^n$ on the interior of each such region. At boundaries between regions, the particles undergoing Brownian motion move with equal probability into each of the neighbouring regions. We prove this gives rise to a well-defined Markov process in Theorem 3.1. The method of proof is adapted from [3], Theorem 3.1, which is an analogous result for the existence of Brownian motion on 2-dimensional Euclidean complexes. Our main result is Theorem 4.6, in which we prove that suitably defined random walks converge to Brownian motion on BHV tree-space. The method of proof involves projecting sample paths in tree-space down onto a single positive orthant $\mathbb{R}^n_+$. The reflection principle can then be used to calculate probabilities for sets of sample paths on tree-space in terms of the probability of the sets of projected paths.

The remainder of the paper is structured as follows. Section 2 contains background material on the geometry of tree-space and Brownian motion in Euclidean space; we prove a result concerning intersections between Brownian sample paths and coordinate hyperplanes which the proof of convergence relies on. In Section 3 we prove that Brownian motion exists as a well-defined Markov process on tree-space, and relate the probability measure on sample paths to the corresponding probability measure for reflected Brownian motion in Euclidean space. In Section 4 we define random walks on BHV tree-space and prove they converge to Brownian motion under a certain limit. When $x_0$ is not a binary tree, both existence of the stochastic processes and convergence of the random walk require special attention, and we do not give complete results for this case. This situation is discussed in Section 5. We give some concluding remarks in Section 6.

2. Background

2.1. The geometry of tree-space. The combinatorial and geometric structure of Billera-Holmes-Vogtmann tree-space was originally described in [2]. We give a brief description here of the essential elements we require, but more detail can be found in the original paper. In particular, we are more concerned with the combinatorial
structure; geodesics and the corresponding metric structure are not required for most of our results.

Tree-space $T_N$ is the set of edge-weighted unrooted trees with $N$ leaves labelled by a 1-to-1 correspondence with a fixed set of species $S = \{1, 2, \ldots, N\}$. The space of rooted trees can easily be obtained from the unrooted space by adding an extra species 0 to $S$, so that 0 is by definition attached to the root vertex. The edge weights take values in $\mathbb{R} > 0$ and are usually represented graphically as edge lengths. Each element of $T_N$ contains $N$ edges which end in a leaf, referred to as pendant edges. Trees are called binary or (from the biological literature) fully-resolved if every vertex has degree 3 apart from the leaves, in which case the tree contains exactly $2N - 3$ edges. A tree containing fewer internal edges is called unresolved, and will contain at least one vertex of degree 4 or more. For convenience we define $N' = N - 3$, the number of internal edges on a fully-resolved tree. Two trees $T_1, T_2$ are said to have the same topology if they are identical when the edge weights are ignored. There are $(2N - 5)!!$ fully-resolved topologies for $N$ species.

A split is a partition of $S$ into two disjoint subsets. Given a tree $x \in T_N$, cutting any edge on $x$ partitions $S$ exactly this way, so edges are determined by the splits they induce. Any tree can therefore be regarded as a weighted set of splits which satisfy a certain compatibility relation. Arbitrary sets of splits do not represent trees: for example two the splits $\{1, 2\}$, $\{3, 4, \ldots, N\}$ and $\{1, 3\}$, $\{2, 4, \ldots, N\}$ are not compatible and cannot be simultaneously represented on a tree. Given a tree $x$ and a split $e$ we write $\ell_x(e)$ to denote the weight (or length) assigned to $e$ in $x$, and define $\ell_x(e) = 0$ when $e$ does not correspond to any edge in $x$. Tree topologies correspond to unweighted sets of compatible splits. We say a fully-resolved topology $\tau$ resolves a topology $\tau'$ when, as a set of splits, $\tau' \subset \tau$. Splits are pendant if they correspond to pendant edges, or internal otherwise.

The weights associated with pendant edges will be ignored for the present. The set of all trees with some fixed fully-resolved topology $\tau$ can then be put into correspondence with the interior of the positive orthant $\mathbb{R}^N_+ = [0, \infty)^N$ by associating each internal split with a coordinate axis in $\mathbb{R}^N$. By taking a tree with topology $\tau$ and shrinking an edge down to length zero, so that the corresponding split is removed from the tree, a point on the boundary of $\mathbb{R}^N_+$ is obtained. The boundary of $\mathbb{R}^N_+$ therefore corresponds to the set of trees with topologies resolved by $\tau$. Specifically, each codimension-$k$ face of $\mathbb{R}^N_+$ corresponds to a topology in which $k$ internal splits have been removed from $\tau$. The orthant $O_\tau \subset T_N$ consists of all trees with topology $\tau$ together with all trees with unresolved topologies $\tau' \subset \tau$, and so $O_\tau \cong \mathbb{R}^{N'}_+$. Suppose a single edge is contracted down to length zero in a fully resolved tree with topology $\tau$, and let $\tau'$ denote the resultant unresolved topology. In particular, $\tau'$ will contain a single vertex of degree 4, and suppose the four subtrees attached to this vertex are denoted $A, B, C, D$. Then there are three topologies which resolve $\tau'$ (including $\tau$): $((A, B), (C, D))$, $((A, C), (B, D))$ and $((A, D), (B, C))$. The process of continuously contracting an edge to give a degree-4 vertex, and then expanding out an alternative edge, is called nearest neighbour interchange (NNI). It follows that each codimension-1 face of an orthant $O_\tau$ corresponds to a set of trees which also lie on a codimension-1 face of exactly two other orthants related to $\tau$ by NNI. Then, ignoring pendant edge weights, tree-space is the union $\bigcup_\tau O_\tau$ where the union is taken over fully resolved topologies, and triplets of orthants related by NNI are glued along codimension-1 faces. The codimension-$k$ faces of the orthants
for $2 \leq k \leq N'$ also overlap and are glued in a similar way. However, it turns out that Brownian motion avoids these faces almost surely, and so the way they are glued together is much less important for our application than the codimension-1 faces. When the weights of pendant edges are introduced, tree-space is the product $\mathbb{R}_{+}^{N} \times \bigcup_{k} \mathcal{O}_{+}$; however, we will continue to ignore pendant edges for the remainder of the paper and take $\mathcal{T}_{N} = \bigcup_{k} \mathcal{O}_{+}$. We let $\mathcal{T}_{N}^{(k)} \subset \mathcal{T}_{N}$ be the set of trees containing at most $N' - k$ internal edges for $k = 1, \ldots, N'$. Equivalently, $\mathcal{T}_{N}^{(k)}$ is the union of the codimension-$k$ faces of the orthants.

It is helpful to consider the structure of tree-space for the cases $N = 4$ and $N = 5$. When $N = 4$ every tree has just one internal edge, and there are three possible fully resolved tree topologies, as specified in the previous paragraph. Thus $\mathcal{T}_{4}$ consists of three copies of $\mathbb{R}_{+}$ glued together at the origin, and we write $\mathcal{T}_{4} = \vee^{3} \mathbb{R}_{+}$. Each ‘arm’ of $\mathcal{T}_{4}$ corresponds to a different topology, and the position along the arm specifies the length of the internal edge. The origin corresponds to the tree with no internal edges, called a star tree. For $N = 5$ there are 15 different fully resolved topologies, each containing 2 internal edges. In this case, $\mathcal{T}_{5}$ consists of 15 copies of $\mathbb{R}_{+}^{2}$ glued in sets of 3 along their edges. The origin of each orthant corresponds to the star tree. This article contains several figures showing a few orthants of $\mathcal{T}_{5}$ laid side-by-side in the plane: see Figure 1 for example.

Billera, Holmes and Vogtmann proved the existence of a unique geodesic between any two points in $\mathcal{T}_{N}$ consisting of straight line segments in each orthant, and such that the length of the geodesic is given by the sum of the $L^{2}$-lengths of the segments. We will refer to the corresponding metric as the geodesic distance on tree-space. Within each orthant it is the same as the Euclidean ($L^{2}$) metric. However, beyond this, the structure and properties of tree-space geodesics are not directly relevant to our application.

Although we ignore pendant edge weights for the remainder of this article, they are very readily dealt with. For each tree $x \in \mathcal{T}_{N}$, the pendant edge weights correspond to a point in $\mathbb{R}_{+}^{N}$. If we consider Brownian motion on this space, with reflection at the boundaries, then at time $t_{0}$ the distribution of particles starting from $x_{0}$ is given by a wrapped multivariate normal density with variance-covariance matrix $t_{0} \times I$. Reflected random walks on $\mathbb{R}_{+}^{N}$ converge to Brownian motion. Existence of Brownian motion and convergence of random walks on the full tree-space $\mathbb{R}_{+}^{N} \times \bigcup_{k} \mathcal{O}_{+}$ therefore follow immediately if we can establish these properties for the internal edges.

2.2. Brownian motion in Euclidean space. Before considering Brownian motion in tree-space, we start by establishing a result in Euclidean space, Lemma 2.1. This will be used later to obtain a convenient partition of the set of Brownian sample paths in tree-space.

Let $X = \mathbb{R}^{N'}$, fix $x_{0} \in X$ and fix a time $t_{0} > 0$. Let $C_{x_{0}}[0, t_{0}](X)$ denote the metric space of maps $\eta : [0, t_{0}] \to X$ continuous with respect to the Euclidean metric $d_{2}(\cdot, \cdot)$ on $X$ which satisfy $\eta(0) = x_{0}$. The metric between paths $\eta_{1}, \eta_{2} \in C_{x_{0}}[0, t_{0}](X)$ is defined by

$$d_{\infty}(\eta_{1}, \eta_{2}) = \sup_{t \in [0, t_{0}]} d_{2}(\eta_{1}(t), \eta_{2}(t)).$$

Then Brownian motion on $X$ for particles starting from $x_{0}$ and of duration $t_{0}$ defines a distribution $B_{x_{0}, t_{0}}(X)$ on $C_{x_{0}}[0, t_{0}](X)$ equipped with the Borel sigma algebra.
Define

$$\Pi_i = \{(x_1, \ldots, x_{N'}) \in X : x_i = 0\}$$

for each $i = 1, \ldots, N'$ and

$$X^{(1)} = \bigcup_i \Pi_i \quad \text{and} \quad X^{(2)} = \bigcup_{i \neq j} (\Pi_i \cap \Pi_j).$$

Any path $\eta$ drawn from $B_{x_0,t_0}(X)$ misses the origin $0 \in X$ almost surely for all $t > 0$. In fact, because sample paths for Brownian motion on $\mathbb{R}^2$ miss the origin almost surely, it follows that any path $\eta$ drawn from $B_{x_0,t_0}(X)$ will almost surely have no intersection with $X^{(2)}$ for $t > 0$.

**Lemma 2.1.** If $x_0 \notin X^{(2)}$ and $\eta$ is drawn from $B_{x_0,t_0}(X)$ there exists almost surely a finite sequence $i_1, i_2, \ldots, i_k \in \{1, 2, \ldots, N'\}$ for some $k \geq 0$ such that the intersections of $\eta$ with $X^{(1)}$ lie on $\Pi_{i_1}, \Pi_{i_2}, \ldots, \Pi_{i_k}$ in turn, ignoring multiple consecutive intersections with the same hyperplane, so that $i_r \neq i_{r+1}$ for all $r$.

**Proof.** Since $\eta$ misses $X^{(2)}$ almost surely, it follows that there is almost surely a well defined, though possibly infinite, set of intersections with $X^{(1)} \setminus X^{(2)}$, ordered according to increasing time. Each such intersection is with a single hyperplane, and the corresponding time-ordered index set of hyperplanes is denoted $\tilde{\pi}(\eta)$. If $\eta(t) \in \Pi_i$ for some $i$ and some $t \in (0, t_0)$, then it may be the case that there is an infinite set $S_i^t \subset (t - \epsilon, t + \epsilon)$ for some $\epsilon > 0$ such that $\eta(s) \in \Pi_i$ for all $s \in S_i^t$ and such that $\eta(s) \cap \Pi_j = \emptyset$ for all $s \in (t - \epsilon, t + \epsilon)$ and $j \neq i$. In other words, given $\eta$ hits a hyperplane $\Pi_i$ there might be an infinite number of other hits with the same hyperplane before the path $\eta$ hits a different hyperplane. Given $\eta$ we define the *pairwise distinct* sequence of hyperplanes $\pi(\eta)$ to be the subsequence of $\tilde{\pi}(\eta)$ obtained by ignoring any such repeated hits with the same hyperplane. Therefore, to prove the lemma it remains to show that $\pi(\eta)$ is finite almost surely. The basic idea is that for $\pi(\eta)$ to be infinite, $\eta$ must hit $X^{(2)}$, and this happens with zero probability. This method of proof is adapted from [3], Theorem 3.1, and we adopt similar notation.

Given a point $x \in X^{(1)} \setminus X^{(2)}$ let $Q(x)$ be the closed ball around $x$ with diameter $\text{dist}(x, X^{(2)})$. Define random times $S_0 = 0$ and $T_i = \inf\{t > S_{i-1} : \eta(t) \in X^{(1)}\}$, $S_i = \inf\{t > T_i : \eta(S_i) \notin Q(\eta(T_i))\}$ for $i = 1, 2, \ldots$. If $\pi(\eta)$ is infinite, then there will be an infinite sequence of pairs $S_i, T_i$, bounded above by $t_0$, so that $S_i - T_i \to 0$. If we assume that the event $\pi(\eta)$ is infinite occurs with positive probability, it follows that, with positive probability, the radius of $Q(\eta(T_i))$ tends to zero and that $\eta$ hits $X^{(2)}$, since the domain $[0, t_0]$ is compact and the continuous function $\text{dist}(\eta(t), X^{(2)})$ achieves its bounds. However, $\eta$ hits $X^{(2)}$ with probability zero, contradicting the assumption. \hfill $\Box$

It is also useful to consider the reflected Brownian motion on the positive orthant $X_+ = \{(x_1, \ldots, x_{N'}) : x_i \geq 0 \text{ for all } i\}$. The positive orthant is equipped with the standard Borel sigma algebra, and we assume that $x_0 \in X_+$. The map

$$(x_1, \ldots, x_{N'}) \mapsto (|x_1|, \ldots, |x_{N'}|)$$

defined on $X$ extends to give the reflection map

$$\mathcal{R} : C_{x_0}[0, t_0](X) \to C_{x_0}[0, t_0](X_+)$$
which is continuous. This then yields a well-defined probability measure $B_{x_0,t_0}(X_+)$ on $C_{x_0}(0,t_0)(X_+)$ defined by

\[(2.1) \quad B_{x_0,t_0}(X_+)(A) = B_{x_0,t_0}(X)(\mathcal{R}^{-1}A)\]

for any measurable set $A$. As for Brownian motion on $X$, for any sample path $\eta$ of the reflected Brownian motion the sequence $\pi(\eta)$ is almost surely well-defined and finite.

3. Brownian motion in tree-space

In this section we (i) give a brief proof of the existence of Brownian motion as a Markov process on tree-space, and (ii) relate the corresponding distribution of sample paths to Brownian motion in $\mathbb{R}^{N'}$. First, however, we briefly consider Brownian motion on $T_4$, as this plays an important role in the more general case.

3.1. Special case: $T_4$. Explicit solutions to the heat equation on $T_4 = \sqrt{3}\mathbb{R}_+$ have been constructed previously [9] and we briefly describe those solutions here. Let $(x,k)$ denote a point in $\sqrt{3}\mathbb{R}_+$ where $x \in \mathbb{R}_+$ and $k \in \{1,2,3\}$ indexes the three ‘arms’. If Brownian motion starts from $(x_0,l)$, then at time $t_0$ the density function is

\[(3.1) \quad f(x,k;x_0,t) = \frac{2}{3}\phi(x;-x_0,t) + \delta_{kl}(\phi(x;x_0,t) - \phi(x;-x_0,t))\]

where $\phi(x;\mu,\sigma^2)$ denotes the density of the normal distribution on $\mathbb{R}$ with mean $\mu$ and variance $\sigma^2$. This is the same as a certain transition density used in Theorem 3.1 of [3] and we make use of it in our proof of existence of Brownian motion on tree-space below. By construction, Brownian motion on each arm is locally the same as Brownian motion on $\mathbb{R}$; and given a particle hits the origin in some time interval $(t_1, t_2)$ then it is equally likely to be on each of the three arms at time $t_2$. If we project $\sqrt{3}\mathbb{R}_+ \to \mathbb{R}_+$, in the obvious way, the density function obtained is a wrapped normal distribution, corresponding to Brownian motion on $\mathbb{R}_+$.

The space $\sqrt{3}\mathbb{R}_+ \times \mathbb{R}^d$ is called the open book. The heat equation is readily solved on this space: the density function at time $t$ is the product of equation (3.1) with the density of a $d$-dimensional isotropic multivariate normal distribution with variance $t$. It has the same properties as Brownian motion on $\sqrt{3}\mathbb{R}_+$: away from $0 \times \mathbb{R}^d$ the process is the same as Brownian motion on $\mathbb{R}^{d+1}$, and given a particle hits $0 \times \mathbb{R}^d$ then it is equally likely to be found on each ‘page’ of the book at future times.

3.2. Existence. Brownian motion in $T_N$ is defined in the following way. Within each orthant, the process is identical to Brownian motion in $X_+ = \mathbb{R}^{N'}_+$. Then, when a diffusing particle hits the codimension-1 boundary of an orthant, it moves with equal probability into each of the three neighbouring orthants related by NNI. Brownian sample paths in $X_+$ hit codimension-2 faces with probability zero, and this essentially ensures the process on $T_N$ is well-defined. Brin and Kifer [3] proved the existence of Brownian motion on 2-dimensional Euclidean complexes as well-defined Markov processes. Their construction and proof can be adapted to $T_N$. We describe the steps involved below.

Our aim is to define in distribution the Markov process $Z(t)$ corresponding to Brownian motion in $T_N$, for all $t \geq 0$. In fact, we restrict to the case $Z(0) \in T_N \setminus T_N^{(2)}$. (The case $Z(0) \in T_N^{(2)}$ is discussed in Section 5.) We need some notation:
suppose that \( x \in T_N^{(1)} \setminus T_N^{(2)} \) so that \( \text{dist}(x, T_N^{(2)}) > 0 \) and \( x \) lies on the boundary of exactly three orthants \( \mathcal{O}_i \subset T_N, \ i = 1, 2, 3 \). Let \( Q_i(x) \subset \mathcal{O}_i \) for \( i = 1, 2, 3 \) be defined by

\[
Q_i(x) = \{ y \in \mathcal{O}_i : |\ell_y(e) - \ell_x(e)| \leq \frac{1}{2} a(x) \text{ for all splits } e \text{ in } \mathcal{O}_i \}
\]

where \( a(x) = \text{dist}(x, T_N^{(2)}) \) and \( \ell_y(e) \) denotes the edge length associated with split \( e \) in tree \( y \). Note that for exactly one split \( e_i^* \) in each \( \mathcal{O}_i \), \( \ell_x(e_i^*) = 0 \) and so the inequality in equation (3.2) becomes \( \ell_y(e_i^*) \leq a(x)/2 \). Let \( Q(x) = \bigcup Q_i(x) \). Then \( Q(x) \) is isomorphic to \( (\sqrt{3}[0, a(x)/2]) \times [0, a(x)]^{N'-1} \), which is a subset of the open book discussed in section 3.1.

For a point \( x_0 \in T_N \setminus T_N^{(1)} \) we define \( Z(t) \) starting from \( x_0 \) as the standard \( N' \)-dimensional Brownian motion in the orthant \( \mathcal{O} \) containing \( x_0 \), until it hits a face of \( \mathcal{O} \). For an initial point \( x_0 \in T_N \setminus T_N^{(1)} \), the process \( Z(t) \) is defined to be the diffusion on the open book \( \sqrt{3}\mathbb{R}_+ \times \mathbb{R}^{N'-1} \) restricted to \( Q(x_0) \) until it exits from \( Q(x_0) \). These rules are then applied in an alternating fashion to construct \( Z(t) \) for all \( t \geq 0 \). Brownian motion in the open book is identical to Brownian motion in \( \mathbb{R}^{N'} \) away from the ‘spine’, as explained above, and so the two rules are identical on the interior of each orthant.

**Theorem 3.1.** When \( Z(0) \notin T_N^{(2)} \) the above construction defines \( Z(t) \) in distribution for all \( t \geq 0 \), and \( Z(t) \) avoids \( T_N^{(2)} \) almost surely.

**Proof.** This is an adaptation of the proof of Theorem 3.1 in [3] to the context of tree-space. Since Brownian motion in \( X_+ \) hits \( X^{(2)} \) with zero probability, it follows from the definition of Brownian motion within each orthant that \( Z(t) \) hits \( T_N^{(2)} \) with zero probability. Next, define random times \( S_0 = 0 \) and \( T_i = \inf\{t > S_{i-1} : Z(t) \in T_N^{(1)} \} \), \( S_i = \inf\{t > T_i : Z(t) \notin Q(Z(T_i))\} \) for \( i = 1, 2, \ldots \). Then the construction above defines \( Z(t) \) for all \( t \) provided \( S_i \to \infty \) as \( i \to \infty \) with probability 1. Assume the converse, that \( S_i \to S_{\infty} < \infty \) with some positive probability. Then, with positive probability, \( S_i - T_i \to 0 \) and the side length \( a(Z(T_i)) = \text{dist}(Z(T_i), T_N^{(2)}) \) of \( Q(Z(T_i)) \) tends to 0 as \( i \to \infty \). Now, for \( k = 1, \ldots, N', T_N^{(k)} \) can be decomposed as

\[
T_N^{(k)} = \bigcup_{\tau \in \text{top}(N'-k)} \mathcal{O}_\tau^{(k)}
\]

where \( \text{top}(N'-k) \) denotes the set of topologies containing \( N'-k \) internal edges and for each such topology \( \tau \), \( \mathcal{O}_\tau^{(k)} \) is the closure of the set of trees with that topology. It is therefore possible to find a subsequence \( T_{\tau_1}, T_{\tau_2}, \ldots \) such that \( Z(T_{\tau_1}) \in \mathcal{O}_1^{(1)} \) for all \( i \) and for a particular tree topology \( \tau_1 \in \text{top}(N'-1) \), since \( Z(T_{\tau_1}) \in T_N^{(1)} \) for all \( i \) but there are only finitely many topologies in \( \text{top}(N'-1) \) to choose from. Given any tree \( x \notin \mathcal{O}_1^{(1)} \), \( d(x, Z(T_{\tau_1})) \) is bounded away from zero, but since \( \text{dist}(Z(T_{\tau_1}), T_N^{(2)}) \to 0 \) there exists a topology \( \tau_2 \in \text{top}(N'-2) \) such that \( \tau_2 \subset \tau_1 \) and \( \text{dist}(Z(T_{\tau_1}), \mathcal{O}_{\tau_2}^{(2)}) \to 0 \). We can then consider the behaviour of \( Z(t) \) on \( \bigcup \mathcal{O}_{\tau*} \subset T_N \) where the union is taken over all fully resolved topologies \( \tau^* \) which resolve \( \tau_2 \). By definition, this region includes all the trees which resolve \( \tau_1 \), and so contains \( T_{\tau_1}, T_{\tau_2}, \ldots \). Within this region, \( \text{dist}(Z(t), \mathcal{O}_{\tau_2}^{(2)}) \) has the same distribution as the radial part of 2-dimensional Brownian motion (the Bessel
process). This is because, within this region, the squared distance from $O_\tau(2)$ is given by the sum of the squared edge lengths of the two edges contracted down by projection onto $O_\tau(2)$. However, the Bessel process tends to the origin with zero probability [5], and this contradicts the assumption that $S_\infty < \infty$ with positive probability. It follows that $S_i \to \infty$ as $i \to \infty$ with probability 1, and that the construction gives a well-defined Markov process $Z(t)$. □

3.3. Brownian sample paths in tree-space. The aim in this section is to relate the distribution of Brownian sample paths on $T_N$ to that on $X_+$. The main idea is that there is a natural way to project paths from $T_N$ to $X_+$ providing the initial point $x_0$ is fully resolved. Paths on $T_N$ in the inverse image under projection of a path on $X_+$ differ from one another in the ‘choice’ of orthant taken when traversing $T_N^{(1)}$. The projection enables us to write down the probability measure for certain sets of sample paths of Brownian motion on tree-space in terms of the probability measure on $X_+$.

In analogy to the notation in Section 2.2, we let $C_{x_0}[0,t_0](T_N)$ be the metric space of continuous maps $\eta : [0,t_0] \to T_N$ which satisfy $\eta(0) = x_0$. The metric is defined by

$$d_\infty(\eta_1, \eta_2) = \sup_{t \in [0,t_0]} d(\eta_1(t), \eta_2(t))$$

for any two $\eta_1, \eta_2 \in C_{x_0}[0,t_0](T_N)$ where $d(\cdot, \cdot)$ denotes the geodesic distance. Fix a fully resolved tree $x_0 \in T_N$ with topology $\tau$. Then, by fixing a particular ordering of the internal edges of the tree, we obtain a corresponding point in the interior of $X_+$. By a slight abuse of notion, we also denote this point by $x_0$. We then define the projection map $P$ on paths in the following way. Let $\hat{C}$ be the subset of $C_{x_0}[0,t_0](T_N)$ consisting of paths which do not meet $T_N^{(2)}$ for all $t \geq 0$ and which

\[ \text{Figure 1. Definition of the map } P \text{ for } T_5. \text{ Left: three of the 15 orthants of } T_5 \text{ are shown together with a path starting at } x_0. \text{ Right: projection of the path into } X_+ = \mathbb{R}_+^2. \text{ Each section of the path in } X_+ \text{ is a reflection of the corresponding section within a single orthant in } T_5. \]
lie in \( T_N \setminus T_N^{(1)} \) at time \( t_0 \). For any path \( \eta \in \hat{C} \) we define a corresponding path \( \mathcal{P}(\eta) \in C_{x_0}[0, t_0](X_+) \) in the following way. Under the correspondence between the orthant \( O_\tau \) containing \( x_0 \) and \( X_+ \), we define \( \mathcal{P}(\eta) \) in the natural way until \( \eta \) first hits a codimension-1 face. At this point a certain edge \( e \) has zero length and the corresponding coordinate of \( \mathcal{P}(\eta) \) is zero. Then \( \mathcal{P}(\eta) \) is continued back into the interior of \( X_+ \), maintaining all the coordinate values corresponding to edges other than \( e \). The remaining coordinate is given by the length of the edge \( e^* \) which replaced \( e \) on the path \( \eta \) (so that \( e^* \) is either \( e \) itself or one of the two edges obtained by NNI of \( e \) in \( \tau \)). This process continues for all \( t \in [0, t_0] \) to define \( \mathcal{P}(\eta) \) and gives a continuous map \( \mathcal{P} : \hat{C} \to C_{x_0}[0, t_0](X_+) \). Figure 1 illustrates how \( \mathcal{P} \) is defined.

Next for any sequence \( i_1, i_2, \ldots, i_k \) define \( C_{i_1, i_2, \ldots, i_k} \) by

\[
C_{i_1, i_2, \ldots, i_k} = \{ \eta \in C_{x_0}[0, t_0](X_+) : \pi(\eta) \text{ is well defined} \}
\]

with \( \pi(\eta) = (i_1, i_2, \ldots, i_k) \) and \( \eta(t_0) \in \text{int}(X_+) \} \).

This definition includes for \( k = 0 \) the set \( C_0 \) of paths which do not meet any face. The results in Section 2.2 show that paths sampled from \( B_{x_0, t_0}(X_+) \) are contained in the union of the \( C_{i_1, i_2, \ldots, i_k} \) almost surely, so

\[
\sum_{k \geq 0} \sum_{i_1, i_2, \ldots, i_k} B_{x_0, t_0}(X_+)(C_{i_1, i_2, \ldots, i_k}) = 1.
\]

The second summation is over all sequences \( i_1, \ldots, i_k \) of integers in \( \{1, \ldots, N\} \) such that no consecutive pair of terms in the sequence are equal. We need to consider the pre-images

\[
\hat{C}_{i_1, i_2, \ldots, i_k} = \mathcal{P}^{-1}(C_{i_1, i_2, \ldots, i_k})
\]

which are all open subsets of \( \hat{C} \). Suppose for each codimension-1 face in \( T_N \) we fix an arbitrary numbering \( 1, 2, 3 \) of the adjacent orthants. Then any path \( \eta \) in \( \hat{C}_{i_1, i_2, \ldots, i_k} \) determines a sequence \( j_1, j_2, \ldots, j_k \in \{1, 2, 3\} \) such that \( j_r \) is the number of the final orthant reached before face \( i_{r+1} \) is hit for \( r = 1, \ldots, k-1 \) and \( j_k \) is the number of the orthant containing \( \eta(t_0) \). We then let \( \hat{C}^{j_1, j_2, \ldots, j_k}_{i_1, i_2, \ldots, i_k} \) be the subset of \( \hat{C}_{i_1, i_2, \ldots, i_k} \) of paths which share the same sequence \( j_1, j_2, \ldots, j_k \) and note that

\[
\hat{C}_{i_1, i_2, \ldots, i_k} = \bigcup_{j_1, j_2, \ldots, j_k} \hat{C}^{j_1, j_2, \ldots, j_k}_{i_1, i_2, \ldots, i_k}
\]

is a disjoint union. It follows from the definition of \( \mathcal{P} \) that each set \( \hat{C}^{j_1, j_2, \ldots, j_k}_{i_1, i_2, \ldots, i_k} \) maps to \( C_{i_1, i_2, \ldots, i_k} \) via

\[
\mathcal{P}(\hat{C}^{j_1, j_2, \ldots, j_k}_{i_1, i_2, \ldots, i_k}) = C_{i_1, i_2, \ldots, i_k}.
\]

We will be particularly concerned with the set of Brownian sample paths which end in some particular region of tree-space. Let \( U \) be any subset of \( T_N \). For any \( k \) and sequences \( i = (i_1, \ldots, i_k) \) and \( j = (j_1, \ldots, j_k) \), let \( \hat{C}_{i}^{j}(U) \) be the subset of \( \hat{C}_{i}^{j} \) consisting of paths \( \eta \) with \( \eta(t_0) \in U \). This set will be empty unless the sequences \( i, j \) determine a sequence of topologies ending in an orthant which intersects \( U \). Similarly, given \( i = (i_1, \ldots, i_k) \) let

\[
\hat{C}_{i}(U) = \bigcup_{j \in \{1, 2, 3\}^k} \hat{C}_{i}^{j}(U).
\]
Figure 2. Example of two paths $\eta_1, \eta_2$ in $T_5$ with $P(\eta_1) = P(\eta_2)$. Both paths lie in the same set $\hat{C}_i^j$ with $k = 2$ but traverse a different set of topologies. Left: four of the 15 orthants of $T_5$. Right: projection of the paths into $X_+ = \mathbb{R}_+^2$. The set of paths would be symmetrized around the vertical axis by adding a third path which reflects back into the top-left orthant.

Lemma 3.2. For any subset $U$ of $T_N$ and any sequences $i = (i_1, \ldots, i_k)$ and $j = (j_1, \ldots, j_k)$

$$P^{-1}P\left(\hat{C}_i^j(U)\right) \cap \hat{C}_i^j = \hat{C}_i^j(U)$$

Proof. It is clear that $P^{-1}P(\hat{C}_i^j(U))$ contains $\hat{C}_i^j(U)$, and $P^{-1}P(\hat{C}_i^j(U)) \subset \hat{C}_i$ since $P(\hat{C}_i^j(U)) \subset C_i$. It remains to show the left hand side of (3.3) contains no elements outside $\hat{C}_i^j(U)$. So suppose $\eta \in \hat{C}_i^j(U)$ and $\eta' \in \hat{C}_i^j$ satisfy $P(\eta) = P(\eta')$, so that $\eta'$ is a general element of the left hand side of (3.3). Since both paths are contained in $\hat{C}_i^j$, it follows that $\eta(t_0)$ and $\eta'(t_0)$ lie in the same orthant and also enter that orthant from the same codimension-1 face of tree-space. Then $P(\eta) = P(\eta')$ implies that $\eta$ and $\eta'$ are identical in the orthant containing time $t_0$, and in particular $\eta(t_0) = \eta'(t_0)$ so $\eta' \in \hat{C}_i^j(U)$. This establishes equation (3.3). Note that it is not necessarily the case that $\eta = \eta'$: the paths can pass through different orthants whenever $P(\eta)$ hits the same face of $X_+$ multiple consecutive times – see Figure 2.

The following lemma, which relates the distribution of sample paths on $T_N$ to the distribution on $X_+$ can be thought of as a tree-space analogue of the reflection principle. Equation (3.5) is fundamental: it is equivalent to saying that if a set of paths is fully symmetric with respect to ‘reflection’ into the neighbouring 3 orthants whenever a path hits a codimension-1 boundary, then the probability of a tree-space Brownian sample path lying in that set is given by the probability of the projected set under Brownian motion on $X_+$.  

\begin{align}
\mathbb{P}_{T_N}(\eta \in C_i^j(U)) &= \mathbb{P}_{X_+}(P(\eta) \in C_i^j(U)) \\
\mathbb{P}_{T_N}(\eta \in C_i^j(U)) &= \frac{1}{3} \mathbb{P}_{X_+}(P(\eta) \in C_i^j(U))
\end{align}
Lemma 3.3. If $U \subset \mathcal{T}_N$ is a Borel set then for any $i = (i_1, \ldots, i_k)$ and $j = (j_1, \ldots, j_k)$, $\mathcal{P}(\hat{C}_i^j(U))$ is Borel in $C_{x_0}[0, t_0](X_+)$, and

\begin{equation}
B_{x_0, t_0}(\mathcal{T}_N)(\hat{C}_i^j(U)) = 3^{-k}B_{x_0, t_0}(X_+)(\mathcal{P}(\hat{C}_i^j(U)))
\end{equation}

where $B_{x_0, t_0}(\mathcal{T}_N)$ denotes the probability measure on $C_{x_0}[0, t_0](\mathcal{T}_N)$ induced by Brownian motion.

Proof. Let $A \subset C_i \subset X_+$ be a measurable set for some $i = (i_1, \ldots, i_k)$. Then consider the set of paths $\mathcal{P}^{-1}(A)$. This set of paths is fully symmetric, in the sense that if it contains some path $\eta$ which hits a codimension-1 face in $\mathcal{T}_N$ at time $t$ then it also contains the paths $\eta'$, $\eta''$ which are identical to $\eta$ up to time $t$ but then move into the other two possible orthants at time $t$ and satisfy $\mathcal{P}(\eta) = \mathcal{P}(\eta') = \mathcal{P}(\eta'')$. It follows from the way Brownian motion is constructed on tree-space that

\begin{equation}
B_{x_0, t_0}(\mathcal{T}_N)(\mathcal{P}^{-1}(A)) = B_{x_0, t_0}(X_+)(A).
\end{equation}

If $\eta$ is sampled from $B_{x_0, t_0}(\mathcal{T}_N)$, and given that $\eta \in \hat{C}_i$ for some $i = (i_1, \ldots, i_k)$, then it is equally likely that $\eta$ is contained in each possible set $\hat{C}_i^j$ where $j \in \{1, 2, 3\}^k$ since the direction taken when crossing codimension-1 faces is selected uniformly at random. Since there are $3^k$ possibilities for $j$, it follows that

\begin{equation}
B_{x_0, t_0}(\mathcal{T}_N)(\mathcal{P}^{-1}(A) \cap \hat{C}_i^j) = 3^{-k}B_{x_0, t_0}(X_+)(A).
\end{equation}

By substituting in $A = \mathcal{P}(\hat{C}_i^j(U))$ and applying Lemma 3.2, we obtain equation (3.4).

We also need to consider measurability of sets of sample paths. First note that if $V \subset X_+$ is a Borel set then so is $C(V) = \{\eta \in C_{x_0}[0, t_0](X_+) : \eta(t_0) \in V\}$ and hence also $C_i(V) = C(V) \cap C_i$. Next, fix some open set $U \subset \mathcal{T}_N$. Then $\hat{C}_i^j(U)$ is open in $C_{x_0}[0, t_0](\mathcal{T}_N)$ and

\begin{equation}
V = \{\eta(t_0) : \eta \in \mathcal{P}(\hat{C}_i^j(U))\}
\end{equation}

is also open. In fact if $\mathcal{O}_\tau$ is the orthant corresponding to the end-points of paths in $\hat{C}_i^j$ then the action of $\mathcal{P}$ determines an identification $\mathcal{O}_\tau \cong X_+$, and $V \cong U \cap \mathcal{O}_\tau$ under this identification. Then $\mathcal{P}(\hat{C}_i^j(U)) = C_i(V)$ which is open. The steps taken to generate the Borel $\sigma$-algebras on $\mathcal{T}_N$ and $X_+$ from open subsets can be applied to the sets of paths ending in those subsets of $\mathcal{T}_N$ and $X_+$. It follows that whenever $U \subset \mathcal{T}_N$ is a Borel set, then so is the set $V \subset X_+$ corresponding to the end-points of $\hat{C}_i^j(U)$, and hence $\mathcal{P}(\hat{C}_i^j(U))$ is a Borel set. \hfill \Box

4. Random walks in tree-space and convergence to Brownian motion

We aim to prove that certain random walks on $\mathcal{T}_N$ converge to Brownian motion on $\mathcal{T}_N$ in an appropriate sense. We do this via random walks on Euclidean space $X$, which we define in the following way. Fix a number of steps $m$ and consider the following algorithm:

Algorithm 4.1. Start with $y_0 = x_0 \in X$. For $j = 1, 2, \ldots, r$ where $r = m \times N'$, repeat the following:

1. Pick $k$ uniformly at random from $\{1, \ldots, N'\}$.
2. Sample a realization $\delta_j$ from $N(0, t_0/m)$ and let $\ell^* = \delta_j + y_{j,k}$ where $y_{j,k}$ is the $k$-th coordinate of $y_j$.
3. Set $y_{j+1} := y_j$ but then change the $k$-th coordinate to $y_{j+1,k} = \ell^*$. 
Other versions of this algorithm could also be used. For example, rather than sampling which edge to change at step 1, a fixed permutation of edges can be used, with each edge incremented exactly \( m \) times at step 2. This algorithm also produces paths which converge to Brownian motion on \( X \).

By linearly interpolating between the collection of points \( y_0, y_1, \ldots, y_r \) produced by the algorithm, so that each segment has time duration \( t_0/r \), we obtain an element of \( C_{x_0}[0, t_0](X) \). We denote the distribution on \( C_{x_0}[0, t_0](X) \) induced by this procedure as \( W_{x_0, t_0}^m(X) \). Standard theory of Euclidean random walks shows that \( W_{x_0, t_0}^m(X) \xrightarrow{w} B_{x_0, t_0}(X) \) as \( m \to \infty \) where \( w \) denotes weak convergence of the probability measures [6]. We obtain the random walk on \( X_+ \) by mapping paths on \( X \) to \( X_+ \) via the reflection map. The corresponding distribution on \( C_{x_0}[0, t_0](X_+) \) is denoted \( W_{x_0, t_0}^m(X_+) \) and is defined by

\[
W_{x_0, t_0}^m(X_+)(A) = W_{x_0, t_0}^m(X)(R^{-1}A)
\]

for any Borel set \( A \). Random walks on \( X_+ \) converge to Brownian motion:

\[
(4.1) \quad W_{x_0, t_0}^m(X_+) \xrightarrow{w} B_{x_0, t_0}(X_+) \text{ as } m \to \infty,
\]

This is a straightforward consequence of the convergence of the processes on \( X \).

Next we define the random walk on \( T_N \) via the following algorithm.

**Algorithm 4.2.** Start with a fully resolved tree \( y_0 = x_0 \in T_N \). We maintain a list \( L \) throughout, with \( L \) initially empty. For \( j = 1, 2, \ldots, r \) where \( r = m \times N' \) repeat the following:

1. Pick a split \( e \) from \( y_{j-1} \) uniformly at random.
2. Sample a realization \( \delta_j \) from \( N(0, t_0/m) \) and let \( \ell^* = \delta_j + \ell_{y_{j-1}}(e) \) where \( \ell_{y_{j-1}}(e) \) is length of \( e \) in \( y_{j-1} \).
3. Set \( y_j := y_{j-1} \) but then change a single edge length as follows:
   (a) If \( \ell^* > 0 \) set \( \ell_{y_j}(e) := \ell^* \).
   (b) Otherwise let \( e^* \) be a split selected uniformly at random from the set \( \{e, e', e''\} \) where \( e', e'' \) are splits associated with the two edges obtained by performing NNI of \( e \) in \( y_{j-1} \). Replace \( e \) with \( e^* \) in \( y_j \) and set \( \ell_{y_j}(e^*) := -\ell^* \). If \( e^* = e \), let \( L := L \cup \{j\} \).

Modifications to the algorithm, which deal with the case that \( x_0 \) is unresolved, are discussed in section 5. As for Algorithm 4.1, there are alternative ways to define random walk on tree-space. For example, a fixed order of edges could be used at step 1 of algorithm 4.2 rather than randomly sampling edges; or a uniform distribution could be used to increment edge lengths at step 2. It is also possible to change all the edge lengths simultaneously by using a multivariate normal distribution, although the algorithm is more complicated and so we do not give the details here.

We can use Algorithm 4.2 to simulate paths \( \eta \in C_{x_0}[0, t_0](T_N) \) by linear interpolation via geodesic segments between the points \( y_0 = x_0, y_1, \ldots, y_r \) generated by the algorithm. This process is illustrated in Figure 3. Each geodesic segment involves a single edge in the tree changing length and possibly being replaced with an NNI alternative. Each geodesics consists of either a straight line within a single orthant, or a straight line consisting of two segments in neighbouring orthants, as Figure 3 shows. The list \( L \) records those steps of the algorithm when the simulated random walk meets a face of an orthant and moves into the same orthant rather than either of the two neighbours. In the case that \( j \in L \), the points \( y_{j-1} \) and \( y_j \)
Figure 3. A realization of a random walk lying on three orthants in $T_5$ with each orthant shown as a shaded square. At each step of the random walk a single edge changes length or is replaced with an alternative. The path between points $a, b, c$ corresponds to a single step of the algorithm: from point $a$, the random walk reflected back from point $b$ to give point $c$ in the original orthant. In terms of Algorithm 4.2, this corresponds to a step for which $\ell^*$ was negative and for which $e^* = e$.

are not joined directly by the geodesic $\gamma(y_{j-1}, y_j)$, but by a path which touches the orthant boundary. Specifically, if $e$ is the edge whose length differs between $y_{j-1}$ and $y_j$, then the path consists of contracting $e$ down to zero length from $y_{j-1}$ and then expanding the edge out to hit $y_j$. In Figure 3 such a step occurs between points $a, b, c$.

**Lemma 4.3.** If $\eta$ is a path generated via Algorithm 4.2 then $\mathcal{P}(\eta)$ has the distribution $W_{x_0, t_0}(X_+)$.  

**Proof.** The vector of edge lengths defined by the algorithm is a random walk on $X_+$, generated in exactly the same way as performing Algorithm 4.1 on $X$ followed by the action of the reflection map $R$.  

Let $W_{x_0, t_0}(T_N)$ denote the distribution on $C_{x_0}[0, t_0](T_N)$ induced by Algorithm 4.2. We will not in fact prove that $W_{x_0, t_0}(T_N) \xrightarrow{w} B_{x_0, t_0}(T_N)$ as $m \to \infty$, but rather prove that

$$W_{x_0, t_0}(T_N)(A) \to B_{x_0, t_0}(T_N)(A)$$

for certain sets $A$ consisting of all paths which end in given regions of $T_N$ at time $t_0$. This result is sufficient to enable us to use Algorithm 4.2 to simulate approximately the distribution of particles in tree-space at time $t_0$ assuming they have undergone Brownian motion starting from $x_0$.

**Definition 4.4.** Let $W(x_0, t_0; m)$ denote the distribution on $T_N$ given by the endpoints $y_r$ simulated by Algorithm 4.2, or equivalently given by sampling a path $\eta$
from \( W^m_{x_0,t_0}(\mathcal{T}_N) \) and taking \( \eta(t_0) \). Similarly let \( B(x_0, t_0) \) be the distribution on \( \mathcal{T}_N \) induced in a similar way, but for which \( \eta \) is drawn from \( B_{x_0, t_0}(\mathcal{T}_N) \). Then

\[
W(x_0, t_0; m)(A) = W^m_{x_0, t_0}(\mathcal{T}_N)(\hat{C}(A)) \\
B(x_0, t_0)(A) = B_{x_0, t_0}(\mathcal{T}_N)(\hat{C}(A))
\]

for any measurable \( A \subset \mathcal{T}_N \), where \( \hat{C}(A) \subset \hat{C} \) denotes the set of all paths which hit \( A \) at time \( t_0 \).

If \( P \) is a measure on a measurable metric space, then a \( P \)-continuity set \( U \) is a Borel set such that \( P(\partial U) = 0 \) where \( \partial U \) is the boundary of \( U \). In order to prove weak convergence, it is sufficient to show that \( W(x_0, t_0; m)(U) \to B(x_0, t_0)(U) \) for any \( B(x_0, t_0) \)-continuity set \( U \).

**Lemma 4.5.** If \( U \subset \mathcal{T}_N \) is a \( B(x_0, t_0) \)-continuity set, then

i. \( \hat{C}^j_i(U) \) is a \( B_{x_0, t_0}(\mathcal{T}_N) \)-continuity set for any sequences \( i, j \), and

ii. \( \mathcal{P}(\hat{C}^j_i(U)) \) is a \( B_{x_0, t_0}(X_+) \)-continuity set.

**Proof.** Fix a \( B(x_0, t_0) \)-continuity set \( U \subset \mathcal{T}_N \) and sequences \( i, j \) of length \( k \). Let \( O_i \) be the orthant corresponding to the end points of paths determined by the sequences \( i, j \). The boundary of \( \hat{C}^j_i(U) \) decomposes into two pieces: \( \hat{C}^j_i(\partial U) \) and a set of paths which end at time \( t_0 \) on \( \mathcal{T}_N^{(1)} \). Both these have zero measure with respect to \( B_{x_0, t_0}(\mathcal{T}_N) \) so:

\[
B_{x_0, t_0}(\mathcal{T}_N)(\partial \hat{C}^j_i(U)) = B_{x_0, t_0}(\mathcal{T}_N)(\hat{C}^j_i(\partial U)) = 0
\]

and \( \hat{C}^j_i(U) \) is a \( B_{x_0, t_0}(\mathcal{T}_N) \)-continuity set. Given \( i, j \), \( \mathcal{P} \) is an isomorphism between \( \hat{C}^j_i \) and \( C_i \). It follows that \( \partial \mathcal{P}(\hat{C}^j_i(U)) \) decomposes into two parts: \( \mathcal{P}(\hat{C}^j_i(\partial U)) \) and another set of paths which end on the boundary of \( X_+ \) and which therefore have zero measure with respect to \( B_{x_0, t_0}(X_+) \). Thus if \( U \) is a \( B(x_0, t_0) \)-continuity set, using Lemma 3.3 we have:

\[
0 = B_{x_0, t_0}(\mathcal{T}_N)(\hat{C}^j_i(\partial U)) \\
= 3^{-k} B_{x_0, t_0}(X_+)(\mathcal{P}(\hat{C}^j_i(\partial U))) \\
= 3^{-k} B_{x_0, t_0}(X_+)(\partial \mathcal{P}(\hat{C}^j_i(U)))
\]

and so \( \mathcal{P}(\hat{C}^j_i(U)) \) is a continuity set with respect to \( B_{x_0, t_0}(X_+) \). \( \square \)

**Theorem 4.6.**

\[
W(x_0, t_0; m) \xrightarrow{w} B(x_0, t_0)
\]

as \( m \to \infty \).

**Proof.** Fix any Borel set \( U \subset \mathcal{T}_N \), and fix \( k \), \( i = (i_1, \ldots, i_k) \) and \( j \in \{1, 2, 3\}^k \). The proof of Lemma 3.3 also applies to random walks, giving

\[
W^m_{x_0, t_0}(\mathcal{T}_N)(\hat{C}^j_i(U)) = 3^{-k} W^m_{x_0, t_0}(X_+)(\mathcal{P}(\hat{C}^j_i(U)))
\]
which is analogous to equation (3.4). If \( U \) is a \( B_{x_0,t_0}(\mathcal{T}_N) \)-continuity set then as \( m \to \infty \)
\[
W_{x_0,t_0}(\mathcal{T}_N)(\hat{C}_i^j(U)) = 3^{-k}W_{x_0,t_0}(X_+)(\mathcal{P}(\hat{C}_i^j(U)))
\to 3^{-k}B_{x_0,t_0}(X_+)(\mathcal{P}(\hat{C}_i^j(U))) = B_{x_0,t_0}(\mathcal{T}_N)(\hat{C}_i^j(U)).
\]
(4.3)

Convergence in equation (4.3) occurs because random walk on \( X_+ \) converges weakly to Brownian motion, and by Lemma 4.5 \( \mathcal{P}(\hat{C}_i^j(U)) \) is a continuity set with respect to \( B_{x_0,t_0}(X_+) \).

Now for any continuity set \( U \)
\[
W(x_0,t_0;m)(U) = \sum_{k=0}^{\infty} \sum_i \sum_j W_{x_0,t_0}^m(\mathcal{T}_N)(\hat{C}_i^j(U)).
\]
Since this is bounded above by 1, for any \( \epsilon > 0 \) there exists \( K \) such that:
\[
| \sum_{k>K} \sum_i \sum_j W_{x_0,t_0}^m(\mathcal{T}_N)(\hat{C}_i^j(U)) | < \frac{\epsilon}{4},
\]
and
\[
| \sum_{k>K} \sum_i \sum_j B_{x_0,t_0}(\mathcal{T}_N)(\hat{C}_i^j(U)) | < \frac{\epsilon}{4}.
\]

Then, by taking \( m \) sufficiently large
\[
| \sum_{k \leq K} \sum_i \sum_j (W_{x_0,t_0}^m(\mathcal{T}_N)(\hat{C}_i^j(U)) - B_{x_0,t_0}(\mathcal{T}_N)(\hat{C}_i^j(U))) | < \frac{\epsilon}{2}
\]
and so \( |W(x_0,t_0;m)(U) - B(x_0,t_0)(U)| < \epsilon \). This proves the weak convergence in Equation (4.2).

5. Unresolved source \( x_0 \)

In Theorem 3.1 it was assumed that \( x_0 \) contains at least \( N' - 1 \) edges, and in Theorem 4.6 it was assumed that \( x_0 \) was fully resolved. In this section we consider existence of Brownian motion and convergence of random walks when these assumptions are dropped. First we need to extend Algorithm 4.2 when \( x_0 \) is unresolved. This is achieved by redefining \( y_0 \) in the following way. A fully resolved topology which resolves \( x_0 \) is chosen uniformly at random. Then, as a set of weighted splits, \( y_0 \) is taken to have this topology. We set \( \ell_{y_0}(e) = \ell_{x_0}(e) \) for all splits \( e \) in \( x_0 \) but set \( \ell_{y_0}(e) = 0 \) for the remaining splits in the fully resolved topology. Algorithm 4.2 then proceeds in the same manner, sampling \( y_1, \ldots, y_r \) in turn. With this definition in place, we turn attention to existence of the Brownian motion.

Theorem 3.1 above proves existence of Brownian motion when \( x_0 \) is fully resolved or when \( x_0 \in \mathcal{T}_N^{(1)} \setminus \mathcal{T}_N^{(2)} \). In the proof of existence of Brownian motion on 2-dimensional Euclidean complexes in [3], from which Theorem 3.1 was adapted, the case that the Brownian motion starts from a vertex \( v \) required special attention: see [3], Theorem 3.2. It is this theorem we need to adapt to prove existence of the Brownian motion for \( x_0 \in \mathcal{T}_N^{(2)} \). The main idea is that there is a weak limit of the Markov process transition probabilities as \( x_0 \to v \). This is shown in [3] by working in a form of polar coordinates around \( v \), to decompose the Brownian
motion into a radial component and a transverse component. The radial component is a Bessel process, and it is shown that for small $\delta$, the exit distribution of the Brownian motion on a $\delta$-neighbourhood of $x_0 = \nu$ is uniform. This motivates the definition of $y_0$ in the previous paragraph: in tree-space the Brownian motion will move away from any unresolved $x_0$ into the interior of an orthant containing $x_0$, with each orthant equally likely. However, it is difficult to extend the analysis in [3], Theorem 3.2, over to tree-space due to the greater number of dimensions, and the different types of singularity in the space. For example, consider the case $N = 6$ and a tree $x_0 \in \mathcal{T}_6^{(2)}$ which contains a single internal edge with positive length. Each such tree lies on an axis shared by several 3-dimensional orthants. The Brownian motion near $x_0$ decomposes as a Bessel process representing the perpendicular distance from the axis; a Brownian motion along the axis; and a stochastic process on a space consisting of one circular arc with angle $\pi/2$ for each orthant containing $x_0$. Thus the geometric decomposition of the Brownian motion near $x_0$ depends on the codimension of the singularity represented by $x_0$. However, when $x_0$ is the star-tree the Brownian motion is tractable: at time $t$ the density on each orthant is proportional to a multivariate normal density at the origin and with variance-covariance matrix $t \times I$. It is evident that the random walk from $x_0$ converges to this distribution.

The proof of Theorem 4.6 relies on the projection map $P$ being well-defined and continuous. In fact, when $x_0$ contains $N' - 1$ internal edges, the definition of $P$ in Section 3.3 does not require any modification, and the proof of Theorem 4.6 still holds. However, when $x_0 \in \mathcal{T}_N^{(2)}$ there is no canonical definition of $P$. It is important to note that, due to non-trivial holonomy in tree-space, there is no continuous map $\mathcal{T}_N \to X_+$ which acts as an isomorphism $\mathcal{O}_x \cong X_+$ on each orthant when $N > 4$. (The map $P$ defined in Section 3.3 operates specifically on paths in tree-space, and so the existence of $P$ when $x_0$ is fully resolved does not contradict the previous statement.) A proof of convergence when $x_0 \in \mathcal{T}_N^{(2)}$ might consider a small $\delta$-neighbourhood of $x_0$. The discussion in the previous paragraph suggests that the Brownian motion will exit this neighbourhood at a fully-resolved tree uniformly at random on the boundary of the neighbourhood, and the proof of Theorem 4.6 would then apply from this point onwards. Clearly, the extension of Theorems 3.1 and 4.6 to the general case of unresolved $x_0$ requires more technical analysis than developed here, and we leave this as an open problem.

6. Concluding remarks

We have introduced Brownian motion on tree-space as a means of constructing a family of distributions $B(x_0, t_0)$. Our aim is to use these distributions to build more sophisticated probability models on tree-space, and perform inference for these models. The simplest model is to consider a set of data points $x_1, \ldots, x_n \in \mathcal{T}_N$ as being independent draws from $B(x_0, t_0)$. Inference algorithms for the parameters $x_0, t_0$ under this model are being developed and will be the subject of future publications. However, these are based on the use of random walks to approximate Brownian motion on tree-space, and so a very important first step has been to establish convergence, as demonstrated in this article. Furthermore, the Brownian motion studied in this article could be generalized in several different ways by analogy with diffusion processes in Euclidean space. For example, it might be possible to construct a diffusion with a non-trivial covariance structure corresponding to a
diffusion with ‘preferred directions’ in tree-space. Similarly, it seems possible to adapt Brownian motion in tree-space in order to define a mean reverting stochastic process, an analogue of the Ornstein-Uhlenbeck process. Inference for models using these stochastic process is likely to rely on forward simulation, and will therefore build on the existence and convergence results established in this article.

References

[1] Bacak, M. Computing medians and means in Hadamard spaces. *SIAM J. Optimiz.* 24, 3 (2014), 1542–1566.

[2] Billera, L., Holmes, S., and Vogtmann, K. Geometry of the space of phylogenetic trees. *Adv. Appl. Math.* 27 (2001), 733–767.

[3] Brin, M., and Kifer, Y. Brownian motion, harmonic functions and hyperbolicity for euclidean complexes. *Math. Z.* 237, 3 (2001), 421–468.

[4] Feragen, A., Owen, M., Petersen, J., Wille, M., Thomsen, L., Dirksen, A., and De Bruijne M. Tree-space statistics and approximations for large-scale analysis of anatomical trees. In 23rd biennial International Conference on Information Processing in Medical Imaging (IPMI) (2013).

[5] Itô, K., and McKean, H. P. *Diffusion processes and their sample paths.* Springer, 1965.

[6] Karatzas, I., and Shreve, S. E. *Brownian motion and stochastic calculus.* Springer, 1988.

[7] Miller, E., Owen, M., and Provan, J. S. Polyhedral computational geometry for averaging metric phylogenetic trees. *Adv. Appl. Math.* 68 (2015), 51–91.

[8] Nye, T. An algorithm for constructing principal geodesics in phylogenetic treespace. *IEEE ACM T. Comput. Bi.* 11, 2 (2014), 304–315.

[9] Nye, T., and White, M. Diffusion on some simple stratified spaces. *J. Math. Imaging Vis.* 50 (2014), 115–125.

[10] Skwerer, S., Bullitt, E., Huckemann, S., Miller, E., Oguz, I., Owen, M., Patrangenaru, V., Provan, S., and Marron, J. Tree-oriented analysis of brain artery structure. *J. Math. Imaging Vis.* 50 (2014), 126–143.

[11] Weyenberg, G., Huggins, P. M., Schardl, C. L., Howe, D. K., and Yoshida, R. KDEtrees: non-parametric estimation of phylogenetic tree distributions. *Bioinformatics* 30, 16 (2014), 2280–2287.

School of Mathematics and Statistics, Newcastle University, UK

E-mail address: tom.nye@ncl.ac.uk
URL: http://www.mas.ncl.ac.uk/~ntmwn/