

Brownian Motion, Bridges, and Bayesian Inference in Phylogenetic Tree Space

Zachary McNulty

November 2025

These notes are based on the paper [Woodman and Nye \[2025\]](#), together with some additional background and motivation.

1 Background

The goal of phylogenetics is to use information about a collection of species to infer their evolutionary history: the shared ancestors and speciation events that produced the species observed today. We are typically interested in both the *topology* of the evolutionary tree, which records the sequence of speciation events, and its *metric structure*, which records the elapsed time between them. Historically, inference relied on morphological features, from skeletal structures to plumage patterns. Today, genetic sequence data play an increasingly predominant role.

Given aligned DNA sequences for a particular gene across several species, one can specify a model of sequence evolution and select the tree with the largest likelihood. For example, the Jukes–Cantor model assumes that each nucleotide in $\{A, C, G, T\}$ changes to any of the other three at the same rate. With genome-scale data, however, different genes can support substantially different tree topologies. We therefore need procedures that aggregate information across gene trees.

The simplest approach is concatenation, which treats all available sequence data as one long alignment and applies maximum likelihood once. Consensus methods instead estimate a tree for each gene and use the resulting collection as input to a second procedure. For example, ASTRAL [Mirarab et al. \[2014\]](#) seeks a species tree that maximizes its quartet agreement with the gene trees:

$$\hat{T} = \arg \max_T \sum_{g \in \mathcal{G}} \text{qscore}(g, T).$$

Although many consensus methods are asymptotically consistent, finite-sample guarantees can be difficult to obtain, and their statistical analysis is often tailored to the particular combinatorial procedure.

Ideally, we would like to adapt familiar tools such as concentration of sample means, principal component analysis, and Bayesian inference. Doing so directly in a discrete space with little geometry is difficult. To see what goes wrong, we begin with Fréchet means and then introduce the geometric model used in the paper.

1.1 Fréchet Means

Recall that in $\mathbb{R}^d$ the sample mean $\bar{x}$ of data points $\{x_i\}_{i=1}^n$ has the variational characterization:

$$\bar{x} = \arg \min_{c \in \mathbb{R}^d} \sum_{i=1}^n \|x_i - c\|_2^2$$

This is just a consequence of the bias-variance decomposition of squared error. Likewise, we can extend this to the population level. Given a probability distribution μ on $\mathbb{R}^d$ we know:

$$\mathbb{E}_{X \sim \mu} X = \arg \min_{c \in \mathbb{R}^d} \mathbb{E}_{X \sim \mu} \|X - c\|_2^2$$

The sample identity is the special case in which μ is the empirical distribution of $\{x_i\}$. The *Fréchet mean* extends the ordinary Euclidean mean to a metric space $(\mathcal{X}, d)$. We define an empirical Fréchet mean by

$$\bar{x} := \arg \min_{c \in \mathcal{X}} \sum_{i=1}^n d(x_i, c)^2$$

and define the population version analogously. Under suitable geometric and moment assumptions, the Fréchet mean exists, is unique, and enjoys concentration properties analogous to those of the Euclidean sample mean.

Theorem 1 (Fréchet means) *Let $(\mathcal{X}, d)$ be a complete CAT(0) space, and suppose that μ has finite second moment. Then the Fréchet mean of μ exists and is unique [Sturm \[2003\]](#). Under additional concentration assumptions on μ , the empirical Fréchet mean of independent samples $X_i \sim \mu$ concentrates around the population Fréchet mean [Brunel and Serres \[2025\]](#).*

1.2 Introduction to BHV space

One common way of defining a metric on the space of trees is to count the number of "local" changes it takes to go from one tree to another. An example of such a change is the *nearest neighbor interchange (NNI)* which permutes the four subtrees adjacent to a given edge in an (unrooted) binary tree. Figure 1 gives an example. In the rooted case, this is equivalent to the standard tree rotations found in the likes of AVL trees.

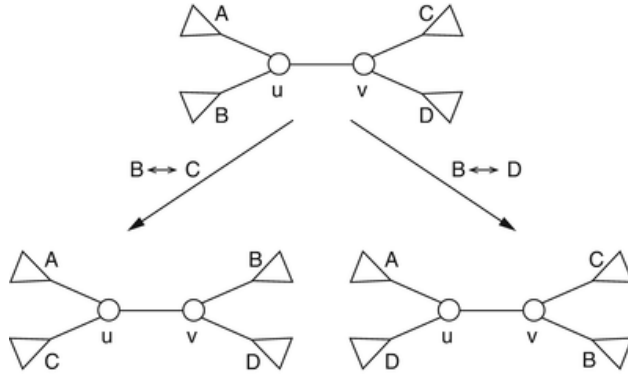


Figure 1: An example of two possible NNI moves.

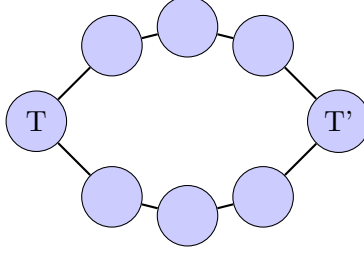


Figure 2: Nonunique paths in a discrete space of phylogenetic trees. Both central vertices are Fréchet means of $\{T, T'\}$.

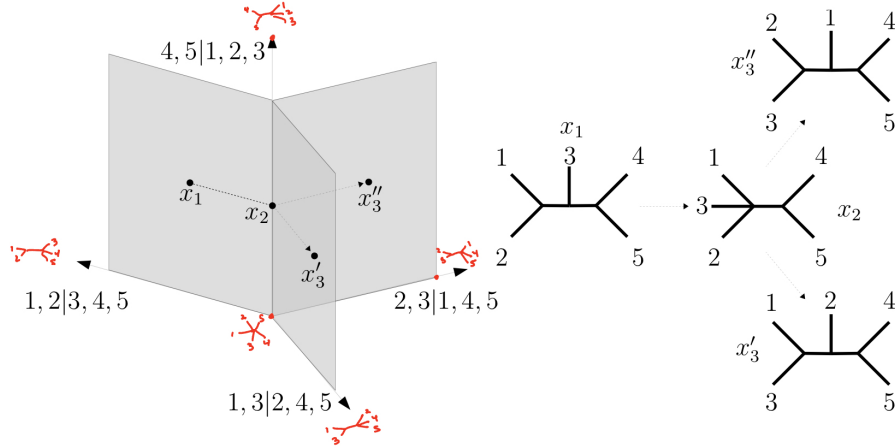


Figure 3: Part of $\mathcal{T}_5$, showing several orthants glued along lower-dimensional faces.

In this picture, the space of phylogenetic trees is just an undirected graph: each vertex is a different tree and an edge connects two trees if they differ by an NNI move. One issue with this space is that shortest paths are not unique.

In such a discrete space, the Fréchet mean need not be unique, and distinct means may be far apart. This is problematic if we want a single estimate $\hat{T}$ of the underlying species tree whose statistical properties can be studied. Billera–Holmes–Vogtmann (BHV) space [Billera et al. \[2001\]](#) was designed in part to address this issue.

Heuristically, the BHV space $\mathcal{T}_N$ of unrooted phylogenetic trees on N labelled leaves is a collection of orthants $\mathbb{R}_{\geq 0}^{N-3}$, one for each fully resolved tree topology. Each coordinate is the length of an internal edge. When an edge length shrinks to zero, the corresponding split disappears; orthants containing that unresolved tree are glued along the resulting lower-dimensional face. Each codimension-one face is incident to three maximal orthants, corresponding to the three resolutions of a degree-four vertex. Part of $\mathcal{T}_5$ is shown in Figure 3.

Theorem 2 *The space $\mathcal{T}_n$ is a CAT(0) space. In particular, it has unique geodesics.*

Consequently, every finite sample has a unique empirical Fréchet mean. If all branch lengths are restricted to a bounded interval, the resulting subset of $\mathcal{T}_N$ is compact, which also facilitates finite-sample concentration around the population Fréchet mean.

2 Brownian Motion in BHV Space

The Fréchet mean $\bar{x}$ gives a model for the "center" of our collection of trees. There are analogous ways of defining its "dispersion". For one, we could look at the sample variance:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n d_{\text{BHV}}(x_i, \bar{x})^2.$$

There is also a substantial literature on central limit theorems for Fréchet means; see, for example, [Hotz et al. \[2022\]](#) and, for stratified spaces in particular, [Mattingly et al. \[2023\]](#).

A more parametric approach assumes that the trees arise from a common model and then estimates its parameters. In Euclidean space, the classical choice would be a Gaussian distribution. There are several inequivalent notions of what a Gaussian should be on a singular metric space; the transition kernel of a Brownian motion provides one natural candidate.

Nye [Nye \[2019\]](#) constructs Brownian motion on cubical complexes; BHV tree space can be treated within this framework after subdividing its orthants into cubes. Let $I = [0, 1]$. A cubical complex is a collection of hypercubes I^n glued along lower-dimensional faces, analogous to a simplicial complex with cubes in place of simplices. It is n -dimensional if it contains an n -cube but no higher-dimensional cube.

Inside an n -cube, B_t should evolve as standard n -dimensional Brownian motion. Upon hitting a codimension-one face f , it chooses uniformly among the n -cubes incident to f and continues there. The construction therefore has two components:

- A reflected Brownian motion in I^n that dictates motion within the current n -cube.
- A discrete random walk on a bipartite graph $G_{\mathcal{X}}$, with one class of vertices representing n -cubes and the other representing $(n-1)$ -faces. An edge records incidence. This walk determines transitions between the n -cubes of $\mathcal{X}$.

The main technical issue is that when we cross f , we will tend to recross it infinitely many times for every positive window of time afterwards before drifting away, so we have to be careful how we define our discrete random walk.

To formalize this picture, first define reflected Brownian motion on I^n by letting its coordinates evolve independently as reflected Brownian motions on $[0, 1]$. Given a continuous path $\eta : [0, 1] \rightarrow I^n$, let $\pi(\eta)$ be the time-ordered sequence of codimension-one faces $(f_1, f_2, \dots)$ visited by η , with consecutive repetitions removed. We must verify that this sequence does not make infinitely many transitions among distinct faces.

Lemma 1 *If $x_0 \in I^n$ does not lie in a codimension-two face, then $\pi(\eta)$ is well-defined and finite almost surely under the law of B_t .*

Thus the discrete walk associated with a continuous Brownian path is almost surely well-defined and corresponds to a simple random walk on the incidence graph $G_{\mathcal{X}}$.

Proof 1 (Proof Outline) *Hitting a codimension-two face requires two coordinates (B_t^i, B_t^j) to hit a specified corner simultaneously. Planar Brownian motion hits any fixed point with probability zero, so $\pi(\eta)$ is well-defined almost surely.*

To prove that $\pi(\eta)$ is almost surely finite, first restrict to paths that remain at least distance ρ from every codimension-two face. Consecutive distinct faces are then separated by a positive distance, and uniform continuity of a Brownian path on a compact time interval rules out

infinitely many such transitions. Finally, let $\rho \downarrow 0$.

Now we define our Brownian motion in terms of our reflected Brownian motion W_t on I^n as:

$$\mathbb{P}(B_t \in A) = \sum_{\omega} p(\omega) \mathbb{P}(W_t \in \mathcal{P}(A \cap \widehat{C}^{\omega})), \quad p(\omega) = \prod_i \frac{1}{\deg(f_i)} \quad (1)$$

where the sum ranges over discrete paths $\omega = (u_0, f_1, u_1, \dots, f_k, u_k)$ with consecutive repeated faces removed, $\widehat{C}^{\omega}$ is the set of continuous paths inducing ω , and $\mathcal{P}$ is the projection into the cubical complex. If W_t has zero drift, covariance $\sigma^2 I_n$, and initial point $W_0 = x_0$, then B_t inherits the source and dispersion parameters (x_0, σ^2) .

2.1 Using Brownian Motion for Inference

We now have a Brownian model parameterized by a source x_0 and dispersion σ^2 . Given sampled trees $x_1, \dots, x_n \in \mathcal{T}_N$, we would like to estimate (x_0, σ^2) . A Bayesian approach places a prior π on these parameters and uses the observed trees to form their posterior distribution.

$$p(x_0, \sigma^2 \mid x_1, \dots, x_n) \propto \pi(x_0, \sigma^2) \prod_{i=1}^n p_{B_1}(x_i \mid x_0, \sigma^2).$$

Here $p_{B_1}(\cdot \mid x_0, \sigma^2)$ denotes the transition density of $B_t(x_0, \sigma^2)$ at time $t = 1$. This density is intractable: Equation (1) gives a construction of the process, but not a readily evaluable transition kernel.

Even normalizing relatively simple distributions is challenging because the number of intersected orthants depends on the center and radius. For example, $\mathcal{T}_4$ is a tripod. The one-dimensional volume of $B_{\text{BHV}}(x_0, \epsilon)$ is

$$3\epsilon - \min\{d_{\text{BHV}}(x_0, 0), \epsilon\},$$

as illustrated in Figure 4.

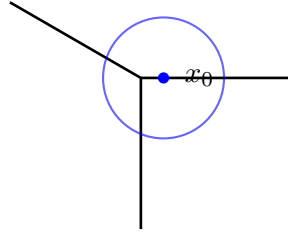


Figure 4: An ϵ -ball in $\mathcal{T}_4$.

This problem is only exacerbated near lower-dimensional faces, where balls intersect many orthants in complicated ways. Thus we approximate B_1 in two steps:

1. We approximate our continuous walk by a discrete walk.
2. If the steps of our discrete walk are small enough, they typically cross at most one face (of codimension-1). In this case, the distribution of the steps is much easier to understand.

2.1.1 Discrete approximation using Gaussian via geodesic firing

Suppose that x_0 lies in the interior of a maximal orthant. Given a dispersion parameter τ^2 , define the *Gaussian via geodesic firing* distribution $GGF(x_0, \tau^2)$ as follows:

- Sample a direction $v \sim N(0, \tau^2 I_{N-3})$.
- Follow the unique geodesic starting at x_0 in direction v .
- Upon reaching a codimension-one face, choose uniformly between the two other maximal orthants incident to that face and continue the unfolded geodesic in the chosen orthant.
- Continue until you travel total distance $|v|$.

If we iterate this process, we obtain an analog of the simple random walk with Gaussian increments. Unsurprisingly, we get the following central limit theorem.

Lemma 2 *Let $y_0 = x_0$ and sample $y_j \mid y_{j-1} \sim GGF(y_{j-1}, \sigma^2/m)$. If $W(x_0, \sigma^2; m)$ denotes the distribution of y_m , then $W(x_0, \sigma^2; m) \Rightarrow B_1(x_0, \sigma^2)$ as $m \rightarrow \infty$.*

Hence our first approximation:

$$p(x_0, \sigma^2 \mid \mathbf{x}) \approx \pi(x_0, \sigma^2) \prod_{i=1}^n p_W(x_i \mid x_0, \sigma^2; m),$$

up to its normalizing constant. The endpoint density p_W remains intractable because it integrates over all compatible random-walk paths. For large m , however, each individual step is short and typically crosses at most one codimension-one face, so its GGF transition density is tractable. This motivates augmenting the model with the intermediate points of each path.

2.1.2 Latent paths and marginalization

For observation x_i , write $y_{i,0} = x_0$, $y_{i,m} = x_i$, and let $\mathbf{y}_i = (y_{i,1}, \dots, y_{i,m-1})$ be the intermediate path. The endpoint density is the integral

$$p_W(x_i \mid x_0, \sigma^2; m) = \int \prod_{j=1}^m f_{GGF}\left(y_{i,j} \mid y_{i,j-1}, \frac{\sigma^2}{m}\right) d\mathbf{y}_i.$$

Rather than evaluate this integral directly, the paper treats the paths as latent variables. The augmented posterior is

$$p(x_0, \sigma^2, \mathbf{y}_1, \dots, \mathbf{y}_n \mid \mathbf{x}) \propto \pi(x_0, \sigma^2) \prod_{i=1}^n \prod_{j=1}^m f_{GGF}\left(y_{i,j} \mid y_{i,j-1}, \frac{\sigma^2}{m}\right).$$

An MCMC algorithm can target this density without evaluating the intractable normalizing constant. Marginalizing the sampled paths then yields posterior samples for (x_0, σ^2) . The remaining challenge is to construct effective proposals for paths conditioned on their two endpoints—that is, bridges.

2.1.3 Sampling bridges

The bridge proposal is inspired by the standard Brownian bridge. If $Y_0, Y_1, \dots, Y_m$ is a Gaussian random walk, then

$$Y_j \mid \{Y_{j-1} = y_{j-1}, Y_m = x^*\} \sim N\left(y_{j-1} + \frac{x^* - y_{j-1}}{m - j + 1}, \frac{m - j}{m - j + 1} \frac{\sigma^2}{m} I_{N-3}\right).$$

Thus the conditional mean lies a proportion $1/(m - j + 1)$ of the way from y_{j-1} to x^* . In BHV space, the analogous basic proposal starts at $y_0 = x_0$, conditions on $y_m = x^*$, and repeats the following for $j = 1, \dots, m - 1$:

1. Construct the geodesic Γ_{y_{j-1}, x^*} and let $\mu_j = \Gamma_{y_{j-1}, x^*}(1/(m - j + 1))$, where the geodesic is parameterized on $[0, 1]$.
2. Sample $y_j \sim GGF(\mu_j, \tau_{j,m})$, where $\tau_{j,m} = \frac{m-j}{m-j+1} \frac{\sigma^2}{m}$.
3. If Γ_{y_{j-1}, y_j} is not simple—equivalently, if it meets a face of codimension greater than one—reject the proposal.

This procedure is a proposal distribution, not an exact draw from the conditional bridge law. It is embedded in a Metropolis–Hastings scheme, whose acceptance correction targets the exact bridge distribution for the discretized random walk.

The paper introduces several refinements to improve acceptance and guarantee adequate proposal support, including mixture proposals and partial bridge updates. It also treats singularities near lower-dimensional faces separately, where the basic proposal is least reliable.

3 Final Remarks

The fitted source and dispersion provide analogues of the mean and variance for a sample of trees. Once these parameters have been learned, the model can also generate new trees for posterior prediction or downstream analysis.

Some natural questions I had after reading the article:

- Under what conditions, if any, does the maximum-likelihood estimate of the Brownian source x_0 coincide with the empirical Fréchet mean? Even when they differ, fixing x_0 at the Fréchet mean could provide a useful approximation when estimating σ^2 .
- Can the construction be extended to a reflected diffusion with nonzero drift μ and anisotropic covariance Σ , rather than $\mu = 0$ and $\Sigma = \sigma^2 I$? Such an extension would require a principled rule for transitions at a face. For example, a drift pointing out of the current cube should make continuation into a neighboring cube more likely than reflection back into the current one.

References

- 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. ISSN 0196-8858. doi: 10.1006/aama.2001.0759. URL <https://www.sciencedirect.com/science/article/pii/S0196885801907596>.
- Victor-Emmanuel Brunel and Jordan Serres. Finite sample bounds for barycenter estimation in geodesic spaces, 2025. URL <https://arxiv.org/abs/2502.14069>.
- Thomas Hotz, Huiling Le, and Andrew T. A. Wood. Central limit theorem for intrinsic fréchet means in smooth compact riemannian manifolds, 2022. URL <https://arxiv.org/abs/2210.17191>.
- Jonathan C. Mattingly, Ezra Miller, and Do Tran. Central limit theorems for fréchet means on stratified spaces, 2023. URL <https://arxiv.org/abs/2311.09455>.
- Siavash Mirarab, Rezwana Reaz, Md Shamsuzzoha Bayzid, Tandy Zimmermann, M Shel Swenson, and Tandy Warnow. Astral: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548, September 2014. doi: 10.1093/bioinformatics/btu462. URL <https://doi.org/10.1093/bioinformatics/btu462>.
- Tom M. W. Nye. Convergence of random walks to brownian motion on cubical complexes, 2019. URL <https://arxiv.org/abs/1508.02906>.
- Karl-Theodor Sturm. Probability measures on metric spaces of nonpositive curvature. In *Heat Kernels and Analysis on Manifolds, Graphs, and Metric Spaces*, volume 338 of *Contemporary Mathematics*, pages 357–390. American Mathematical Society, Providence, RI, 2003.
- William M. Woodman and Tom M. W. Nye. Brownian motion, bridges and bayesian inference in phylogenetic tree space, 2025. URL <https://arxiv.org/abs/2506.22135>.