

Variational Phylogenetic Inference in Hyperbolic Space

August 26, 2021

1 Abstract

Hyperbolic spaces allow quality embeddings of nested data structures, such as trees. Recent efforts to embed phylogenetic trees using distance-based optimisation demonstrate promising results for embedding single trees. However, expressing uncertainty in phylogenetics involves a probability distribution over a super-exponential number of trees. Here, using low dimensional hyperbolic tree embeddings, we explore the Bayesian posterior distribution of trees in a continuous manner. First, we empirically demonstrate that a posterior surface can be well approximated with tree embeddings using an MCMC in the embedding space. We also evaluate the potential of variational inference in the embedding space.

2 Introduction

Bayesian phylogenetics has struggled when provided with many taxa. The super-exponential number of combinations of tree topologies is largely responsible for this; even a modest one hundred taxa provides more tree topologies than the number of atoms in the universe, clearly marginalising over these topologies is intractable. Markov Chain Monte Carlo (MCMC) has become a routine way forwards in this high dimensional space. However, its computational performance is lacklustre compared to online viral outbreaks, calling for faster methods.

One possibility is a variational approximation to the posterior, which minimises the divergence between a chosen approximating function and the posterior. This technique avoids the computational burden of computing the marginal likelihood of the data by instead optimising a lower bound for the evidence (ELBO). Furthermore, it is possible to encode multiple tree topologies in a continuous manner using embeddings. Embedding trees into Riemannian manifolds can enable the efficiencies gains from gradient based optimisation, even as tree topologies

change.

A successful embedding requires an isometry between distances in trees and distances in the embedding space. Previous work has shown that ultrametric trees always have an isometric embedding in Euclidean space [?]. Furthermore, there is an isometric embedding in Euclidean space for any rooted phylogenetic tree by taking the square-root of distances between leaves on the tree [de Vienne et al., 2011]. However, Euclidean embeddings generally have a minimum embedding dimension [?], trading one high-dimensional problem for another. Spurred on by quality embeddings of nested data structures (such as trees) in low dimensions [Sala et al., 2018], several recent works turn to hyperbolic space for embeddings [Chami et al., 2020, Wilson, 2021, Matsumoto et al., 2020]. However, none of these works consider the problem of Bayesian inference in hyperbolic space.

In this work, we embed points onto a sphere S^{d-1} in hyperbolic space \mathbb{H}^d , with each point corresponding to a genetic sequence. Initially points are embedded according to their pairwise genetic distance using an approximate strain minimisation algorithm. Each point is equipped with a variational distribution on the sphere. To sample a tree, we draw one point from each distribution and form the neighbour joining tree from their pairwise distance in hyperbolic space. We then optimise these distributions by maximising the ELBO.

3 Hyperbolic Embeddings

The Poincaré ball $\mathbb{P}^d = \{x \in \mathbb{R}^d : \|x\| < 1\}$ models hyperbolic space \mathbb{H}^d using the metric:

$$d(x, y) = \operatorname{arccosh} \left(1 + 2 \frac{\|x - y\|_2^2}{(1 - \|x\|_2^2)(1 - \|y\|_2^2)} \right),$$

where $\|x\|_2$ is the l^2 -norm in \mathbb{R}^d .

We form continuous embeddings of trees using one point in the Poincaré ball $\mathbf{x} \in \mathbb{P}^d$ for each node in the tree. For an unrooted tree with S taxa, there will

be $m = 2S - 2$ nodes locations $\mathbf{X} \in (\mathbb{P}^d)^m$, where $\mathbf{X} = \{\mathbf{x}_i : i = 1, \dots, m\}$. Embedded nodes are connected to form a minimum spanning tree (MST) protocol that ensures internal nodes have three neighbours and tip nodes have one neighbour. Both the branch lengths and tree topology may freely change as nodes move. Once a tree $T(\mathbf{X})$ is formed, its prior probability $p(T)$ and the likelihood of a sequence alignment $p(D|T)$ under a given model and data D may be easily determined.

This method is quite distinct from cost functions that are based on pair-wise distances. For example Chami’s variant on Dasgupta’s cost, the log-a-like used by Wilson or see refs in Chami. Here, nodes placements do not directly contribute to the cost function (tree posterior), they only impact how a tree is formed.

Theorem 1. *There exists an isometric embedding of a finite metric space with $==$ into hyperbolic space.*

Practically, we find a slight improvement of the variational approximation when the embedding is restricted to a sphere. We hypothesise that ultrametric trees can be embedded into hyperbolic space with their leaves lying on a sphere.

4 Sampling Trees in Hyperbolic Space

It is convenient to sample in Euclidean space using common distributions, so we use a homeomorphism between \mathbb{P}^d and \mathbb{R}^d to embed points. The embedding function $g : \mathbb{R}^d \rightarrow \mathbb{P}^d$ is:

$$g(\mathbf{x}) = \frac{\mathbf{x}}{1 + \|\mathbf{x}\|_2}$$

with gradient

$$\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} = \frac{1}{1 + \|\mathbf{x}\|_2} \left(I - \frac{\mathbf{x} \otimes \mathbf{x}}{(1 + \|\mathbf{x}\|_2)\|\mathbf{x}\|_2} \right).$$

Leaves are restricted to a sphere by normalising their radius to a single value. Proposals are still multivariate normals in \mathbb{R}^d , but then the radius is normalised to radius of the first leaf.

We must show that the Markov Chains are ergodic, that is, as $m \rightarrow \infty$ the distribution $p(\mathbf{X}_m|\mathbf{X}_0)$ converges to the posterior distribution regardless of the choice of $p(\mathbf{X}_0)$. If the chain is not ergodic, then the choice of proposal partitions the state space into subsets which cannot be reached from each other. A

point set $\mathbf{X}_m \in \mathbb{R}^d$ can be reached via our normal distribution proposal, however it remains to be seen if this embedding spans all trees. We must know theorem 1.

5 MCMC

Since each set of node embeddings corresponds to a tree which has a well defined likelihood and prior probability, MCMC can proceed with the standard Metropolis-Hastings algorithm. The set of nodes are initialised to locations \mathbf{X}_0 . For each node $\mathbf{x}_i \in \mathbf{X}_0$ a new location is proposed and accepted or rejected according to the Metropolis-Hastings algorithm, giving the next iteration \mathbf{X}_1 .

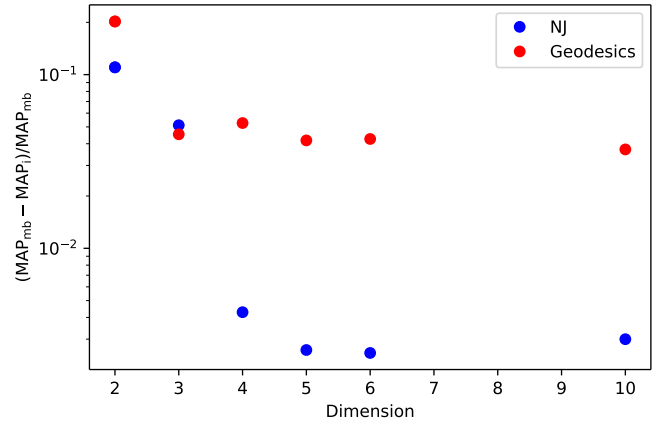


Figure 1: Improvement of the MAP estimate on a tree with 17 taxa with increasing dimension.

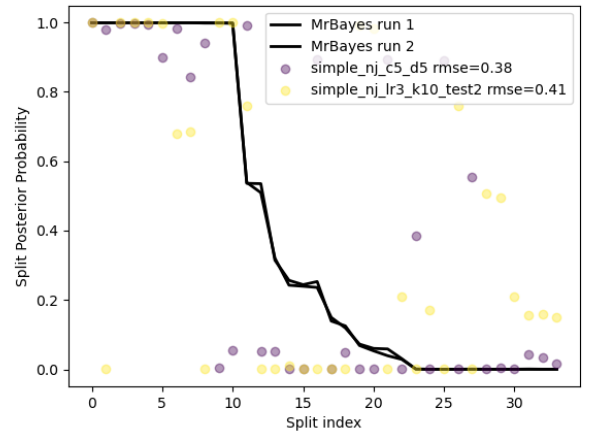


Figure 2: Comparison of splits on a tree with 17 taxa.

6 Variational Inference

Intro to VI==

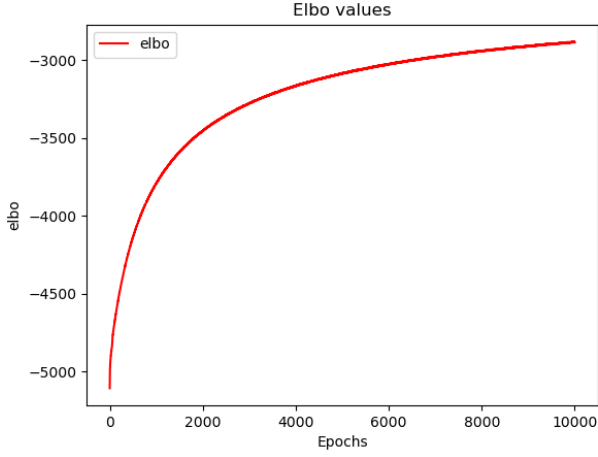


Figure 3: ELBO tree on a tree with 6 taxa. With 17 taxa, the ELBO either decreases or stays roughly constant.

7 Comparison to MrBayes

A posterior was approximated for a 17 taxa set using Dodonaphy’s MCMC and its VI before being compared to MrBayes. A tree was simulated using a birth (rate 2) death (rate .5) model and a sequence alignment was generated from this tree under the JC69 model of genetic evolution. The pairwise patristic distances were computed between the tips on the simulated tree. The tips were initialised in the Poincare ball using hydra. Then the internal nodes were randomly placed in \mathbb{P}^d with uniform directional and radius from a scaled Beta distribution $r \sim s \times \text{Beta}(a = 2, \beta = 5)$ using scales $s \in [0, 2 * \min(d(0, \mathbf{x}_i)]$, where i only includes the tip nodes. Internal node locations were sampled 10^4 times and the initialisation with the highest tree likelihood was selected.

8 Thoughts

8.1 Embedding

- Want a MST with given tips and degree constraints: Given tips see this [stackexchange](#) question. But degree constrained is harder. This algorithm looks at only constraining the .

For $n-2$ iterations:

```
Find the pair of edges i and j \
    neighbouring a single internal \
    vertex that minimises the cost \
    C(i)+C(j)
```

```
Add edges i and j to the tree
```

- I tried without the Matsumoto adjustment and the likelihood was a bit worse.

8.2 MCMC

- What type of tree rearrangements occur when topology changes?
- Posterior surface is smooth if topology doesn't change.
- Remove isometric component of new vector. Want rotations and translations component
- I tried using log-a-like function instead of likelihood [Wilson, 2021] (leaves constrained to a sphere). It gave terrible trees on a six taxa dataset - likelihood ~ -8000 (not -2600) and bad splits. I tried weighting the edges inversely with their length (shorter edges are weighted more) and the result was about the same.
- Try on more taxa 20, 100, 200 using hpc
- How easy is it to add taxa? Might only need pair-wise distance to a subset of other taxa.

8.3 VI

- Pytorch's optimisers don't converge for non-convex problems.
- Posterior "surface" is non-convex and not continuous...
- What if we learn the curvature?

8.4 Full rank

Intuitively, nodes that are close together should be a bit correlated. However, it doesn't seem to improve things much. In the off-diagonals in the covariance matrix are initialised to zero, the ELBO gets much

higher faster compared to if the off-diagonal terms have a non-zero covariance. That said, I haven't run simulations long enough to be sure, only 1000 epochs with a small learning rate of 0.01.

8.5 Distance-based

Could we adopt an approach like Wilson and Chami, where only the distributions of the embedded points are optimised based on their pair-wise distances. Only then do we infer a tree. The advantage of this is that the cost function is differentiable and more in line with what other people in the ML community do. However, this isn't actually modelling the Bayesian Posterior, just a proxy for it.

References

- [Billera et al., 2001] Billera, L. J., Holmes, S. P., and Vogtmann, K. (2001). Geometry of the Space of Phylogenetic Trees. *Advances in Applied Mathematics*, 27(4):733–767.
- [Bose et al., 2020] Bose, A. J., Smofsky, A., Liao, R., Panangaden, P., and Hamilton, W. L. (2020). Latent Variable Modelling with Hyperbolic Normalizing Flows. *arXiv:2002.06336 [cs, stat]*. arXiv: 2002.06336.
- [Chami et al., 2020] Chami, I., Gu, A., Chatzifratris, V., and Ré, C. (2020). From Trees to Continuous Embeddings and Back: Hyperbolic Hierarchical Clustering.
- [Dasgupta, 2016] Dasgupta, S. (2016). A cost function for similarity-based hierarchical clustering. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, STOC '16, pages 118–127, New York, NY, USA. Association for Computing Machinery.
- [de Vienne et al., 2011] de Vienne, D. M., Aguilera, G., and Ollier, S. (2011). Euclidean Nature of Phylogenetic Distance Matrices. *Systematic Biology*, 60(6):826–832.
- [Dinh et al., 2017] Dinh, V., Bilge, A., Zhang, C., and Matsen, F. A. (2017). Probabilistic Path Hamiltonian Monte Carlo. volume 70 of *Proceedings of Machine Learning Research*, page 10, International Convention Centre, Sydney, Australia. PMLR.
- [Greenberg et al., 2020] Greenberg, C. S., Macaluso, S., Monath, N., Lee, J.-A., Flaherty, P., Cranmer, K., McGregor, A., and

- McCallum, A. (2020). Data Structures & Algorithms for Exact Inference in Hierarchical Clustering. *arXiv:2002.11661 [physics, stat]*. tex.ids= greenberg2020dataa arXiv: 2002.11661.
- [Gu et al., 2018] Gu, A., Sala, F., Gunel, B., and Ré, C. (2018). Learning Mixed-Curvature Representations in Product Spaces.
- [Iuchi et al., 2021] Iuchi, H., Matsutani, T., Yamada, K., Iwano, N., Sumi, S., Hosoda, S., Zhao, S., Fukunaga, T., and Hamada, M. (2021). Representation learning applications in biological sequence analysis. *bioRxiv*, page 2021.02.26.433129. Publisher: Cold Spring Harbor Laboratory Section: New Results.
- [Katherine, 2016] Katherine, S. J. (2016). Review Paper: The Shape of Phylogenetic Treespace. *Systematic Biology*, page syw025.
- [Keller-Ressel and Nargang, 2019] Keller-Ressel, M. and Nargang, S. (2019). Hydra: A method for strain-minimizing hyperbolic embedding of network- and distance-based data. *arXiv:1903.08977 [cs, math, stat]*. arXiv: 1903.08977.
- [Layer and Rhodes, 2017] Layer, M. and Rhodes, J. A. (2017). Phylogenetic trees and Euclidean embeddings. *Journal of Mathematical Biology*, 74(1-2):99–111. tex.ids= layer2017phylogenetica.
- [Matsumoto et al., 2020] Matsumoto, H., Mimori, T., and Fukunaga, T. (2020). Novel metric for hyperbolic phylogenetic tree embeddings. preprint, Bioinformatics.
- [Monath et al., 2019] Monath, N., Zaheer, M., Silva, D., McCallum, A., and Ahmed, A. (2019). Gradient-based Hierarchical Clustering using Continuous Representations of Trees in Hyperbolic Space. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 714–722, Anchorage AK USA. ACM.
- [Nagano et al., 2019] Nagano, Y., Yamaguchi, S., Fujita, Y., and Koyama, M. (2019). A Wrapped Normal Distribution on Hyperbolic Space for Gradient-Based Learning. In *International Conference on Machine Learning*, pages 4693–4702. PMLR. tex.ids= nagano2019wrappeda ISSN: 2640-3498.
- [Nye, 2011] Nye, T. M. W. (2011). Principal components analysis in the space of phylogenetic trees. *Annals of Statistics*, 39(5):2716–2739. Publisher: Institute of Mathematical Statistics.
- [Sala et al., 2018] Sala, F., Sa, C. D., Gu, A., and Re, C. (2018). Representation Tradeoffs for Hyperbolic Embeddings. In *International Conference on Machine Learning*, pages 4460–4469. PMLR. ISSN: 2640-3498.
- [Sarkar, 2012] Sarkar, R. (2012). Low Distortion Delaunay Embedding of Trees in Hyperbolic Plane. In van Kreveld, M. and Speckmann, B., editors, *Graph Drawing*, Lecture Notes in Computer Science, pages 355–366, Berlin, Heidelberg. Springer.
- [Sumner, 2017] Sumner, J. G. (2017). Dimensional Reduction for the General Markov Model on Phylogenetic Trees. *Bulletin of Mathematical Biology; New York*, 79(3):619–634. Num Pages: 619-634 Place: New York, Netherlands, New York Publisher: Springer Nature B.V.
- [Whidden and Matsen, 2015] Whidden, C. and Matsen, F. A. (2015). Quantifying MCMC Exploration of Phylogenetic Tree Space. *Systematic Biology*, 64(3):472–491. tex.ids: whidden2015quantifyinga publisher: Oxford Academic.
- [Wilson, 2021] Wilson, B. (2021). Learning phylogenetic trees as hyperbolic point configurations. *arXiv:2104.11430 [cs]*. arXiv: 2104.11430.

9 Appendix

9.1 Numerical stability in Hyperbolic space

As either $\|x\|_2^2 \rightarrow 1$ or $\|y\|_2^2 \rightarrow 1$, eq. 3 can become a numerically unstable way to compute distances. Since the Poincaré ball is a stereographic projection of the hyperboloid, an equivalent metric comes from projecting of x and y into the hyperboloid model of \mathbb{H}^d and using its metric. The hyperboloid model is the sheet inside $\mathbf{x} \in \mathbb{R}^{d+1}$ such that $x_0^2 - \sum_i x_i^2 = 1$. It has metric $d_{hyp}(\mathbf{x}, \mathbf{y}) = \text{acosh}(-\langle \mathbf{x}, \mathbf{y} \rangle)$, where the Lorentz inner product of \mathbf{x} and \mathbf{y} is:

$$\langle \mathbf{x}, \mathbf{y} \rangle = -x_0 y_0 + \sum_{i>0} x_i y_i$$

The stereographic projection onto the hyperboloid $\phi : \mathbb{P}^d \rightarrow \mathbb{H}^d$ takes a point to $\phi(\mathbf{x}) = \left(\frac{(1+\|x\|_2^2)}{(1-\|x\|_2^2)}, \frac{2\mathbf{x}}{(1-\|x\|_2^2)} \right)$. Thus, for $\mathbf{x}, \mathbf{y} \in \mathbb{P}^d$ we can use $d(\mathbf{x}, \mathbf{y}) = d_{hyp}(\psi(\mathbf{x}), \psi(\mathbf{y}))$.

9.2 Normalising Jacobian

Normalising the leaf positions to radius r by $n_r(x) = r\mathbf{x}/\|\mathbf{x}\|$ has Jacobian

$$\frac{\partial n_r(\mathbf{x})}{\partial \mathbf{x}} = r \frac{\partial n_1(\mathbf{x})}{\partial \mathbf{x}} = \frac{r}{\|\mathbf{x}\|} \left(I - \frac{\mathbf{x} \otimes \mathbf{x}}{\|\mathbf{x}\|^2} \right)$$

Wrapping Method We propose a node's new location from a Gaussian by projecting the point from the Poincaré ball into $\Psi : \mathbb{P}^d \rightarrow \mathbb{R}^d$. First project onto the hyperboloid model in \mathbb{R}^{d+1} using $\text{proj}^{-1}(\mathbf{x})$ (see appendix), then projecting onto the tangent plane of $\mathbf{x} = 0$ using the d -dimensional vector $u = (0, 1, 1, 1, \dots)$: $\Psi(\mathbf{x}_i) = \text{proj}^{-1}(\mathbf{x}_i) \cdot u$. A new point is drawn from a Gaussian located at $\Psi(\mathbf{x}_i)$ with given standard deviation, before being projected back to the Poincaré ball with Φ^{-1} .

The determinate of the Jacobian of the projection $\text{proj}_\mu : \exp_\mu \circ \text{PT}\mu_0 \rightarrow \mu$:

$$\left| \frac{\text{proj}_\mu(\mathbf{x})}{\partial \mathbf{x}} \right| = \left(\frac{\sinh(\|\mathbf{x}\|)}{\|\mathbf{x}\|} \right)^{n-1}.$$

The Jacobian of stereographic projection projection back into the Poincaré ball ϕ^{-1} is:

$$\frac{\partial \phi^{-1}(\mathbf{x})}{\partial \mathbf{x}} = \begin{cases} \frac{1}{1+x_0} & \text{if } j = i+1 \\ \frac{-x_{i+1}}{(1+x_0)^2} & \text{if } j = 0 \\ 0 & \text{otherwise} \end{cases}$$

Since this projection is from \mathbb{R}^{d+1} to \mathbb{R}^d , the Jacobian matrix is not square and so we use the d -dimensional Jacobian $|J_{\phi^{-1}}(\mathbf{x})| = (\det(J_{\phi^{-1}}(\mathbf{x})J_{\phi^{-1}}(\mathbf{x})^T))^{1/2}$. The determinant of the Jacobian simplifies to:

$$\left| \frac{\partial \phi^{-1}(\mathbf{x})}{\partial \mathbf{x}} \right| = \frac{(1+x_0)^2 + \sum_{i=1}^d x_i^2}{(1+x_0)^{2(d+1)}}.$$

Discontinuities Note that the likelihood function is discontinuous as the topology changes. This means the optima found may be only locally optimal and may depend on the starting location. The initialisation by Hydra aims to mitigate this effect.

Brute force MST The figure below shows a grid search of the posterior landscape under a MST protocol. We could do a similar thing for neighbour joining now that we're primarily using it.

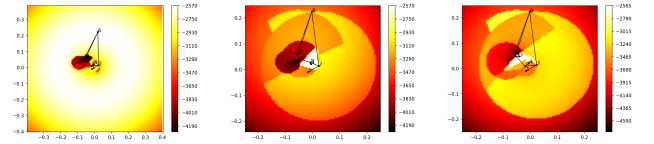


Figure 4: Fix all node positions but one (a) node 6, (b) node 8 and (c) node 9. Move this one node throughout the Poincaré disk and plot the tree posterior by placing the node at that point.