Variational Phylogenetic Inference in Hyperbolic Space

September 9, 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 corrresponding 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 Trees in Hyperbolic Space

3.1 Hyperbolic Space

The most common model of Hyperbolic space is the Poincaré ball $\mathbb{P}^d = \{x \in \mathbb{R}^d : ||x|| < 1\}$ contain points in a unit ball equipped with the metric

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

where $||x||_2$ is the l^2-norm in $\mathbb{R}^d$. However, as either $||x||_2^2 \to 1$ or $||y||_2^2 \to 1$, eq. 3.1 can be-

come a numerically unstable way to compute distances [Sala et al., 2018]. Fortunately, the poincare ball is a stereographic projection of the hyperboloid model, so an equivalent metric comes from projecting of x and y into the hyperboloid model of $\mathbb{H}^d$ and using its metric. The stereographic projection $\phi : \mathbb{P}^d \to \mathbb{H}^d$ takes a point to

$$\phi(\mathbf{x}) = \Big(\frac{(1 + ||x||_2)}{(1 - ||x||_2)}, \frac{2\mathbf{x}}{(1 - ||x||_2}\Big). \tag{1}$$

Then the hyperboloid model $\mathbb{H}^d$ 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

$$d_{hyp}(\mathbf{x}, \mathbf{y}) = \text{acosh}(-\langle \mathbf{x}, \mathbf{y} \rangle), \tag{2}$$

where the Lorentz inner product of $\mathbf{x}$ and $\mathbf{y}$ is:

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

Thus, for $\mathbf{x}, \mathbf{y} \in \mathbb{P}^d$ we can use that $d(\mathbf{x}, \mathbf{y}) = d_{hyp}(\phi(\mathbf{x}), \phi(\mathbf{y}))$. Because of their connection we use these two models somewhat interchangably, modelling points in the Poincaré ball, but computing their distances on the hyperboloid.

3.2 Sampling in Hyperbolic Space

Given a point in hyperbolic space, we want to sample nearby points according to a distribution. To do this, we project points from the Poincaré ball into $\mathbb{R}^d$ and sample distributions there before projecting back into $\mathbb{P}^d$. This enables sampling to use common and efficient sampling techniques in $\mathbb{R}^d$, here we choose a multivariate normal distribution. We consider two possible homeomorphisms $\mathbb{H}^n \to \mathbb{R}^d$: a method of wrapping Euclidean vectors onto the hyperboloid and a homeomorphism from the unit ball into $\mathbb{R}^n$.

3.2.1 Wrapping Distributions

This recently developed method by [Nagano et al., 2019] centres around the projection of a vector $\text{proj}(\mathbf{x})$ from $\mathbb{R}^d$ onto the hyperboloid model, which we then take into the Poincaré ball. The projection is the composition of two fuctions, firstly the vector is mapped via the parallel transport function along the hyperboloid in $\mathbb{R}^{d+1}$ to a desired location μ_0. Then the exponential map wraps the vector in $\mathbb{R}^{d+1}$ on the the hypderboloid $\mathbb{H}^d$:

$$\text{proj}_\mu : \exp_\mu \circ \text{PT}\mu_0 \to \mu. \tag{3}$$

TODO: improve notation === Further details are provided in the appendix.

To project a point from the Poincaré ball to Euclidean space, we first take the inverse stereographic projection from $\mathbb{P}^d$ to $\mathbb{H}^d$ then project "downwards" onto the tangent plane of $\mathbf{x} = 0$ using the d-dimensional vector $u = (0, 1, 1, ...) : \mu = \phi^{-1}(\mathbf{x}_i) \cdot u$. We may then sample a vector v using any distribution in $\mathbb{R}^d$, such as a Gaussian $G(\mu, \Sigma)$, before projecting back to the Poincare ball with $\phi \circ \text{proj}_{\mu)}(v)$.

The determinate of the Jacobian of the projection $\text{proj}_\mu(x)$:

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

The Jacobian of stereographic projection projection back into the poincare 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 $|J_{\phi^{-1}}(\mathbf{x})| = (\det(J_{\phi^{-1}}(\mathbf{x}) J_{\phi^{-1}}(\mathbf{x})^T))^{1/2}$, which simplifies to:

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

3.2.2 A simple homeomorphism

Alternatively, we also consider the following embedding function $g : \mathbb{R}^d \to \mathbb{P}^d$:

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

with gradient

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

Whilst this mapping does not preserve distances it has fewer computational steps.

3.2.3 Forming a Tree

From a set of leaf locations x in the Poincaré ball $x_i \in \mathbb{P}^d$ we seek to construct a tree. One way to form a tree is to use the hyperbolic pairwise distances between leaves and construct the neighbour joining (NJ) tree. Neighbour joining ===

Recently, [Chami et al., 2020] ===

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

TODO: I'm not sure I can derive this. I think most ML work in hyperbolic space just assumes it's an approximate embedding. ===

We hypothesise that ultrametric trees can be embedded into hyperbolic space with their leaves lying on a sphere.

4 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 x^0. For each node x_i^t at step t, a new location is proposed and accepted or rejected according to the Metropolis-Hastings algorithm, giving the next iteration x_i^{t+1}. Node locations are proposed using their projection into $\mathbb{R}^d$, a Gaussian sample is taken before taking its inverse projection back into $\mathbb{P}^d$. For simplicity, leaves are restricted to a sphere by normalising their radius to a single value. That is, each proposal is a Gaussian in $\mathbb{R}^d$, but we then scale it to have the same radius as the first leaf.

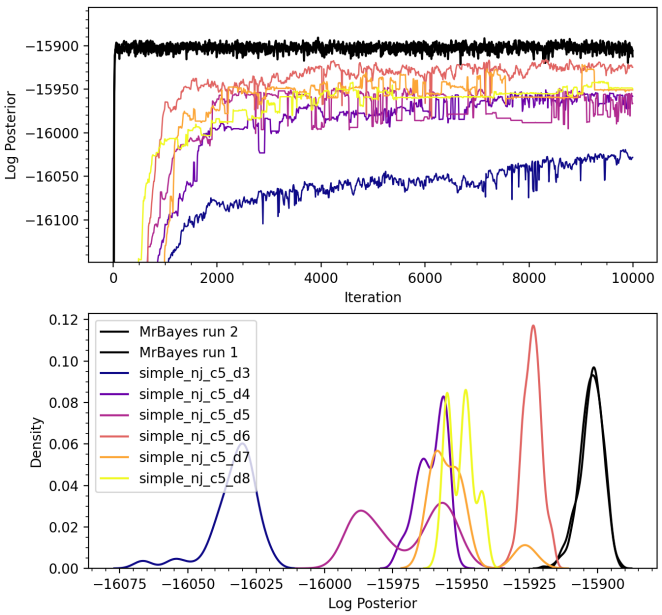

Figure 1: Effect of embedding dimension on MCMC trace. (a) MCMC trace plots of the posterior using the simple embedding method connecting leaves using neighbour joining. (b) Kernel density estimates of the posterior for the last 200 tree in figure (a).

5 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. ===

5.1 Embedding Dimension

Every phylogenetic tree has a Euclidean embedding [de Vienne et al., 2011], however the number

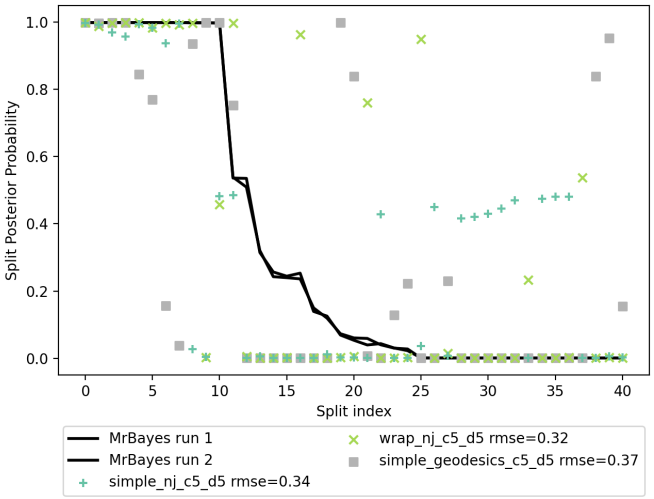

Figure 2: Comparison of splits on a tree with 17 taxa. The two solid lines indicate the posterior frequency of each split (indexed from zero) in two MrBayes runs. Markers represent the posterior frequency of these splits using MCMC in hyperbolic space.

of dimensions m to be sure of an embedding required grows linearly with the number of taxa S [**?**]: $m = S - 1$. On the other hand, hyperbolic embeddings offer low dimensional approximate embeddings of distance spaces. Figure 5.1 empirically demonstrates that the embedding quality saturates after only 4 or 5 dimensions for a tree with 17 taxa.

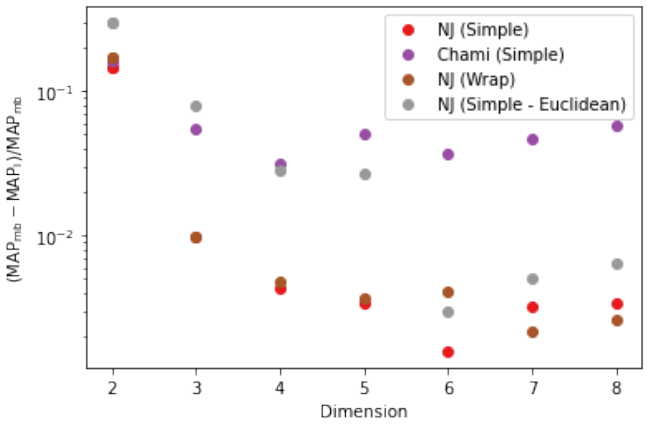

Figure 3: Improvement of the MAP estimate on a tree with 17 taxa with increasing dimension.

6 Variational Inference

In VI, the posterior surface (here a set of surfaces — one for each embedded node i) is approximated by a variational distribution $q_i(z_i)$, each parameterised by parameters z_i. When the data D is fixed, minimising the KL divergence to the posterior $\mathrm{KL}(\mathbf{q}(\phi)||p(T|D))$ is equivalent to maximising the evidence lower bound (ELBO):

$$\mathrm{ELBO}(q) = \mathbb{E}_q[\log(p(z, D))] - \mathbb{E}_q[\log(q(z))].$$

Working with the ELBO is computationally advantageous since it does not depend on the intractable marginal distribution of the data $p(D)$, which is involves integrate over all tree topologies and all branch lengths.

6.1 Differentiable cost

Tree topologies are discrete objects and thus gradient-based operations on embedded trees may be non-differentiable as the topology changes. Finding a way to embed trees such that the likelihood is continuous as the topology changes could lead to local optimisation solutions. However, finding such an embedding remains an open question.

To circumvent this issue, we optimise the likelihood not of a tree, but based on the pair-wise distance of every pair of nodes. Given a set of leaf nodes in the embedding space $x \in \mathbb{H}^d$, the cost function employed is:

$$C = \sum_{i \neq j} w_{ij}(\log L(d_{ij}|D)) + \log p(d_{ij})) \qquad (4)$$

where $L(d_{ij}|D)$ is the likelihood of the distance from x_i to x_j given the sequence alignment data D and the weight w_{ij} applied to each edge sums to one $\sum_{ij} w_{ij} = 1$. The prior probability of the data comes from an exponential prior on the branch lengths. This definition does not requires a time reversible likelihood function, however we use a simple Jukes-Cantor model of evolution which gives $L(d_{ij}) = L(d_{ji})$.

To better mimic the likelihood of a tree, we weight edges according to the Q matrix, which is used to determine which edges are joined during neighbour joining (NJ):

$$Q_{ij} = (S-2)d_{ij} - \sum_{k}^{S} d(i,k) - \sum_{k}^{S} d(j,k).$$

In the algorithm, the most negative entry in Q is selected to and these nodes are connected creating a new ancestor, before again selecting the most negative entry of Q. Accordingly, the weight applied to each edge is constructed as the negative of Q, then a normalised exponential (softmax) function is applied so the weights sum to one:

$$w_{ij} = \frac{\exp(-Q_{ij})}{\sum_{ij} \exp(-Q_{ij})}.$$

6.2 Variational Distributions

We approximated the posterior surface for each leaf node using a log-normal distribution $\mathrm{LogNormal}(x_i, \Sigma_i)$ in Euclidean space $x_i \in \mathbb{R}^d$ wrapped onto a sphere. Specifically, points were drawn from $X \sim \mathrm{LogNormal}(x_i, \Sigma_i)$ then normalised to a lie on a sphere $||X||_2 = ||x||_i$, before being mapped into $\mathbb{H}^d$ using either the wrapping or "simple" embedding methods previously outlined. We assume mean field inference where each Σ_i is independent of Σ_j. Further each Σ_i is a $d \times d$ covariance matrix that we assume is diagonal.

6.3 Implementation

To optimise the ELBO we use stochastic gradient ascent implemented in PyTorch. The autograd functionality of PyTorch automatically computes the gradient of the elbo in order to stochastically optimise with gradients.

6.4 Sampling Trees

At the end of the optimisation, trees must be sampled to approximate the posterior distribution on trees. We sample trees by sampling one point from each variational distribution (one per leaf node) and connect them using neighbour joining. Connecting them in this way does not yield the same ===

Theorem 2. *Could the distribution optimised by cost C equal the neighbour joining tree in the limit?*

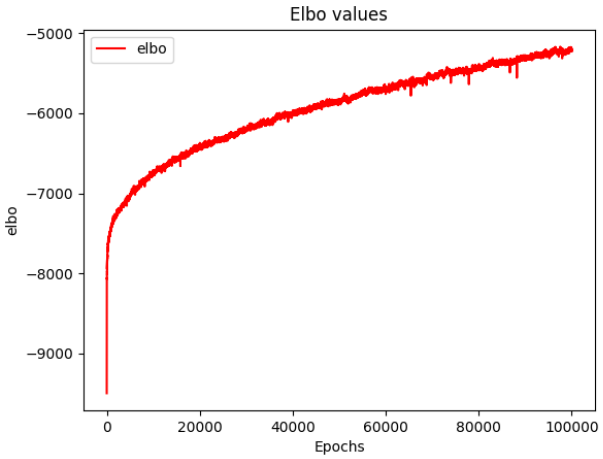

Figure 4: ELBO tree on a tree with 17 taxa in five dimensions. The learning rate is 10^{-1} and there were $k = 2$ importance samples for each ELBO calculation.

7 Acknowledgement

Computational facilities were provided by the UTS eResearch High Performance Computer Cluster.

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., Chatziafratis, 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., Aguileta, 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 hy-

perbolic 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.

8 Thoughts

8.1 Embedding

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

8.2 MCMC

- What type of tree rearrangements occur when topology changes?
- Posterior surfance 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.

9 Appendix

9.1 Normalising Jacobian

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

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

9.2 Details of wrapping

TODO ===

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.

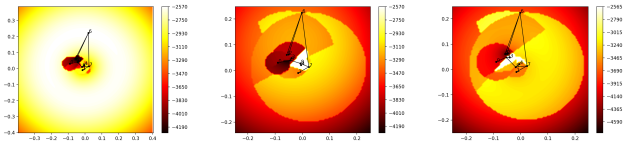

Figure 5: Fix all node positions but one (a) node 6, (b) node 8 and (c) node 9. Move this one node throughout the Poincaré disk and plot the tree posterior by placing the node at that point.