Dodonaphy Notes

July 1, 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 Hyperbolic Space

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

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

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

3 Hyperbolic Embeddings

We form continuous embeddings of trees using one point in the Poincare 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 $\mathbf{X} = \{\mathbf{x}_i : i = 1, ..., m\}$. Embedded nodes are connected to form the 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 aligmnent $\mathrm{p(D}\big|\mathrm{T)}$ 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. The point of using $\mathbb{H}^n$ is to exploit its geometry as a way to move between "adjacent" trees.

Not every tree may be accessible under this MST protocol and in this sense, we have a variational distribution over trees. For example, under a MST it would not be possible to join two very distant edges joined directly if there is an available intermediary node. This raises an interesting question: given a connection protocol, is there a limit to the trees that can be generated?

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 $\mathbf{X}_0$. 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 $\mathbf{X}_1$.

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

4.1 Continuity

Under a MST protocol, the total length of the tree is continuous as nodes move. As a node $\mathbf{x}_i$ moves continuously:

- Case (1) the tree does not change topology. The branch lengths of node i change continuously and so the total length ℓ:

$$\frac{\partial \ell}{\partial \mathbf{x_i}} = cst.$$

- Case (2) the tree changes topology. Suppose x_i moves through the point a and the topology changes from τ_1 to τ_2. The limit from the first topology $\lim_{x_i \to a^{\tau_1}} \ell = \lim_{x_i \to a^{\tau_2}} \ell$.

If the distance between every pair of two nodes $d(i,j)$ accurately reflects the number of mutations per site ν_{ij} between the two, then under a JC69 model, the total branch length ℓ uniquely determines the likelihood function $p(D|T)$.

$$P(\nu_{ij}) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{\nu_{ij}} & i \neq j \\ \frac{1}{4} - \frac{1}{4}e^{\nu_{ij}} & i = j \end{cases}$$

The difficulty is to get the pairwise distances to match the number of mutations per site.

5 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 $\mathrm{KL}(\mathbf{q}(\phi)||p(T\big|D))$ 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. We consider two mean-field variational distributions for the node locations in $\mathbb{P}^d$: a logit-Normal distribution and a "wrapped" Normal distribution as proposed in [**?**].

In both these approaches, the variational distribution $\mathbf{q}$ is smooth, however the posterior probability is not guaranteed to be continuous let alone differentiable as the topology changes. However this difficulty is partly overcome since the ELBO is not a point estimate, but is an expectation taken from k samples. Any discontinuities in $p(z,D)$ are smoothed through sampling under q.

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.

Since the ELBO is an expectation under q, which in this case is a vector of normal distributions $g(x)$,

$$\begin{aligned} \mathrm{ELBO}(q) &= \mathbb{E}_q[\log(p(z,D))] - \mathbb{E}_q[\log(q(z))] \\ &= \int_{\mathbb{R}^d} [\log(p(z,D))] d\mathbf{q} - \int_{\mathbb{R}^d} [\log(q(z))] d\mathbf{q} \\ &= \int_{\mathbb{R}^d} [\log(p(z,D))] \circ g(z) dz - \int_{\mathbb{R}^d} [\log(q(z))] \circ g(z) dz. \end{aligned}$$

Since g(z) is infinitely differentiable and

$$\frac{d}{dz_i}(f \circ g) = f \circ \frac{dg}{dz_i},$$

we find that the elbo is also infinitely differentiable.

6 Results

As a toy example, a posterior was approximated for a six taxa set using Dodonaphy's MCMC and its VI before being compared to BEAST2 — an established MCMC software. 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.

6.1 MCMC

The final log likelihood from BEAST was -6280 (similar to it's log posterior of -6271). In comparison, Dodonaphy MCMC gave -6280, which is a close match. However, figure 1a illustrates that samples from Dodonaphy's MCMC always gave one incorrect split. This could indicate that the MCMC chain is stuck in a local optima and that techniques like using Metropolis-coupled MCMC to escape this optima should be employed.

Figure **??** shows a kernel density estimate for each node location at various stages. Comparing figures **??**a and b reveals that the posterior surface of nodes can drift over the simulation. Indeed tree reconstructions are invariant to isometries on the embedded points since they are only based on their relative distances. Additionally, tree reconstructions won't change as a tip node moves along the surface with constant distance to its nearest neighbour (provided the tree doesn't change topology). The embedded posterior surfaces generated from samples $\mathbf{X}_i$ shouldn't be considered static, but transient over the simulation. Nonetheless, these degrees of freedom should not affect the result of the MCMC.

These densities are clearly not normally distributed. The feasibility of VI rests on how well a simple distribution q can approximate these densities. Boosting to a mixture model may increase the ability for VI to represent these embeddings.

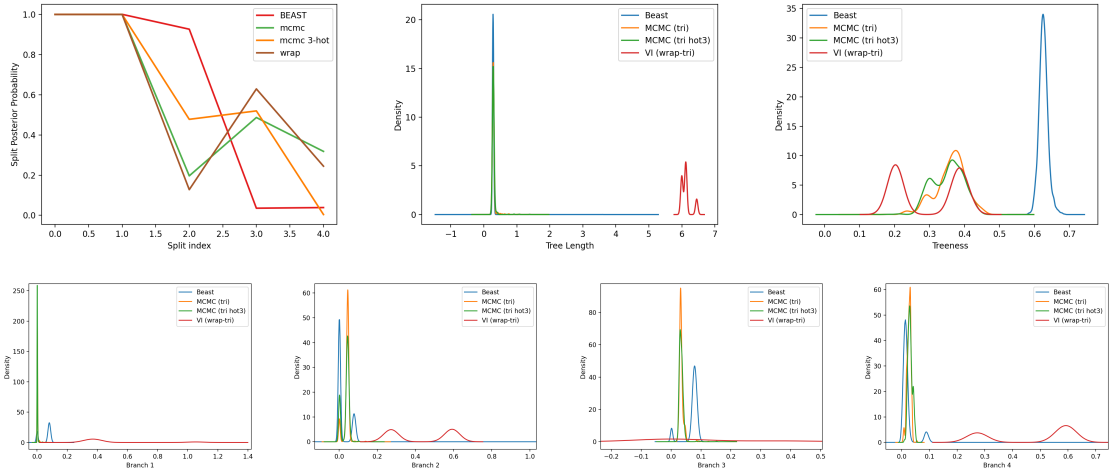

Figure 1: Comparison of tree statistics between Beast and Dodonaphy: (a) split posterior probability (b) total branch length (c) "treeness" (signal/signal+noise) (d-g) branch lengths (possibly not same branch??).

6.2 VI

The traces in figure 2 illustrate that the elbos has a strange start, getting worse for a few hundred epochs, but then increases steadily but probably hasn't converged after the 1e4 epochs.

7 Thoughts

7.1 MCMC

- Try other (Chami) method of connecting?
- What type of tree rearrangements commonly occur?
- Is likelihood under JC69 continuous when topology changes?
- If the nodes move without changing the topology, is the posterior probability smooth?
- Hot chain/ MCMCMC?

7.2 VI

- What if we learn the curvature?
- Effect of $k_{samples}$, boosts, learning rate?

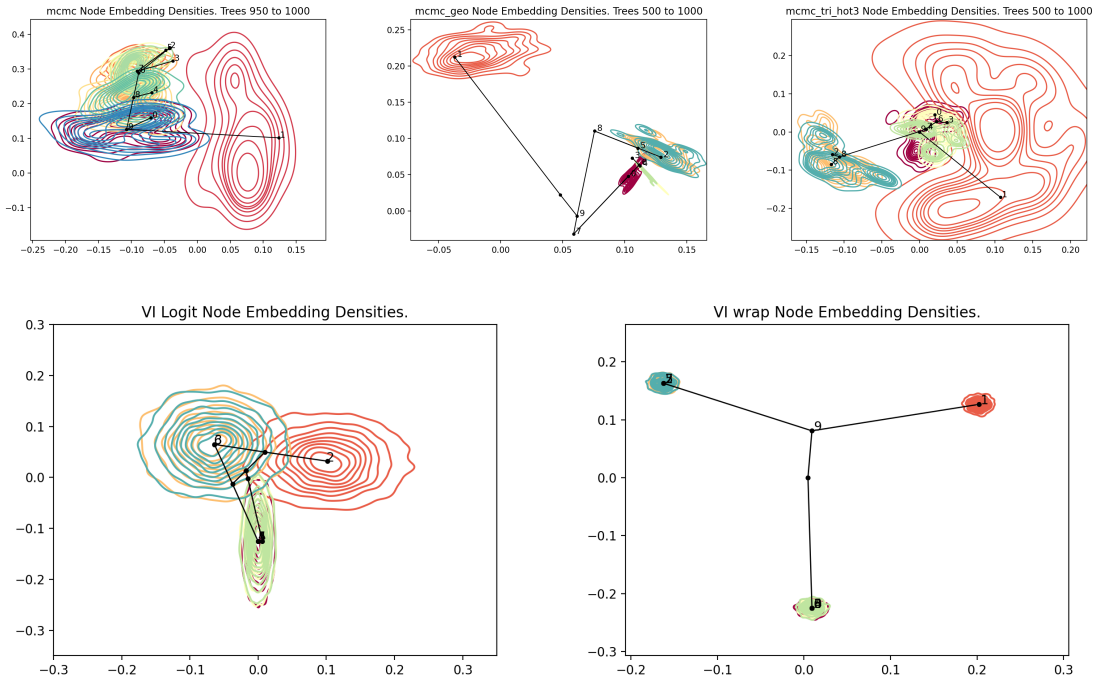

Figure 2: Density estimates of the node locations from Dodonaphy's MCMC projected into $\mathbb{R}^2$. MCMC: using (a) mst, (b) geodesics, (c) incentres . After learning for 10^3 epochs variational densities in $\mathbb{P}^2$ using (d) geodesics, (e) incentres.

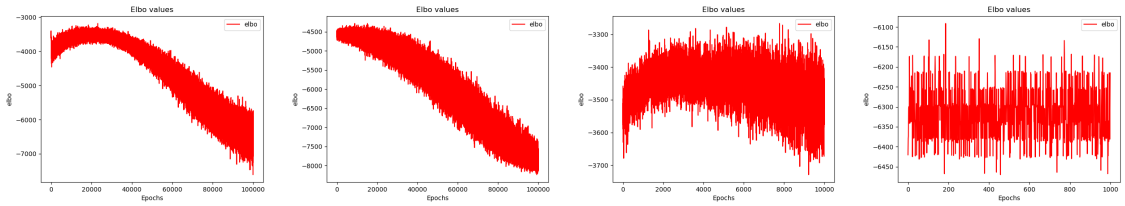

Figure 3: ELBO trace from VI on six taxa using (a) MST + sigmoid transform, (b) MST + wrapping method, (c) geodesics + wrapping, (d) incentres + wrapping.

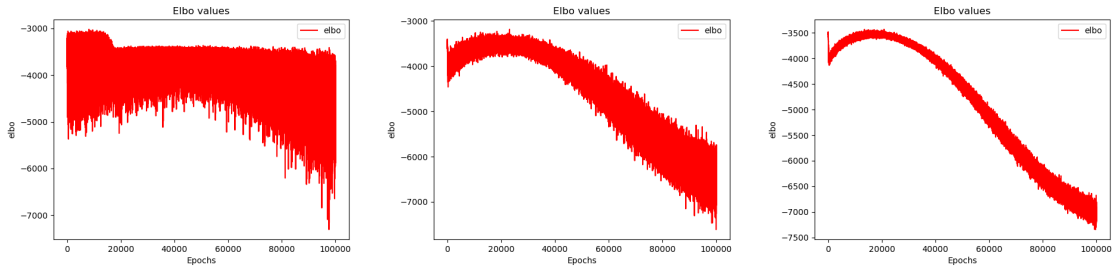

Figure 4: Effect of number of ELBO samples $k = \{1, 10, 100\}$. All using a logit embedding.

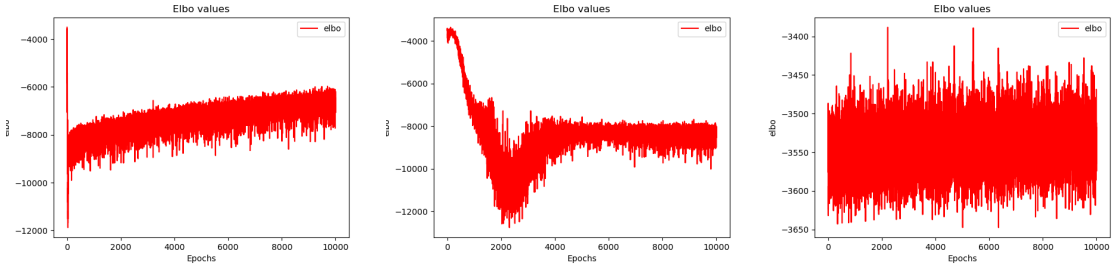

Figure 5: Effect of learning rate $\{1, 10^{-1}, 10^{-4}\}$.

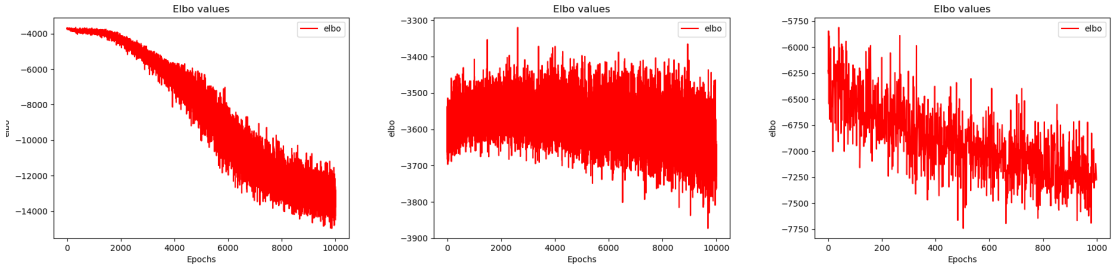

Figure 6: Trying (a) SGD instead of Adam. (b) 3-dimensions. (c) Boosting with a mixture of 3 Gaussians.

- How easy is it to add taxa? Might only need pair-wise distance to a subset of other taxa.
- We could use the original parameterisation with the radius in $r \sim P(N(\mu, \sigma))$ a logit-normal distribution and the directional similarly in a n-dimensional logit-normal.

7.3 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.

7.4 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.

Tips are labelled, internal nodes are not.

8 Appendix

8.1 Numerical stability in Hyperbolic space

As either $||x||_2^2 \to 1$ or $||y||_2^2 \to 1$, eq. 2 can become a numerically unstable way to compute distances. Since the poincare 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 $\phi : \mathbb{P}^d \to \mathbb{H}^d$ takes a point to $\psi(\mathbf{x}) = \left(\frac{(1+||x||_2)}{(1-||x||_2)}, \frac{2\mathbf{x}}{(1-||x||_2}\right)$. Thus, for $\mathbf{x}, \mathbf{y} \in \mathbb{P}_d$ we can use $d(\mathbf{x}, \mathbf{y}) = d_{hyp}(\psi(\mathbf{x}), \psi(\mathbf{y}))$.