Geometry of the space of phylogenetic trees and their limiting behaviors

Jisu KIM

Carnege Mellon University

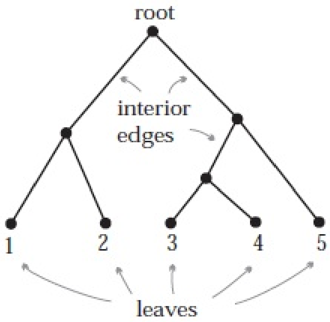
Jan 25, 2017

A metric tree is a labeled tree with lengths on interior edges.

Lemma

([Billera et al., 2001])

- *An m-tree is a tree(a connected graph with no circuits) with a root(distinguished vertex), and m leaves(vertices of degree* 1*), labeled from* 1 *to m.*
- *An edge is interior if it is not connected to a leaf.*
- *A metric m-tree is an m-tree with positive lengths on all interior edges.*



We place an edge directly above the root with the corresponding leaf labeled with 0.

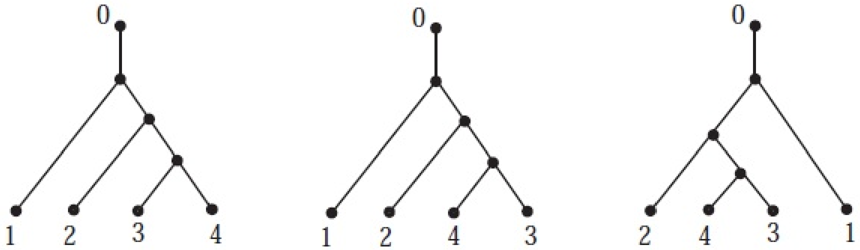


Figure 6: Three pictures of the same tree

The same tree can be embedded differently.

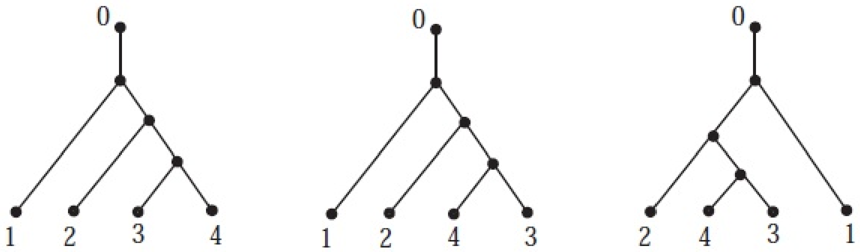


Figure 6: Three pictures of the same tree

Two trees sharing the same combinatorial structure but having leaves labeled differently can be different.

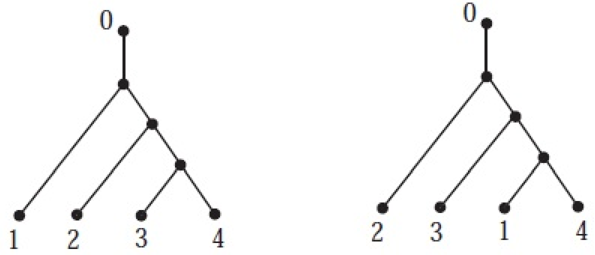


Figure 7: Different trees

Trees with the same combinatorics form an orthant.

- Consider a tree T, with interior edges $e_1, \cdots, e_r$ of lengths $l_1, \cdots, l_r$ respectively. The vector $(l_1, \cdots, l_r)$ specifies a point in the positive open orthant $(0, \infty)^r$.
- Points on the boundary of the orthant (length vectors with at least one coordinate equal to zero) correspond to metric $n-$trees which are obtained from T by shrinking some interior edges of T to 0.
- Each point$\in [0, \infty)^r$ corresponds to a unique metric n-tree.

Trees with the same combinatorics form an orthant.

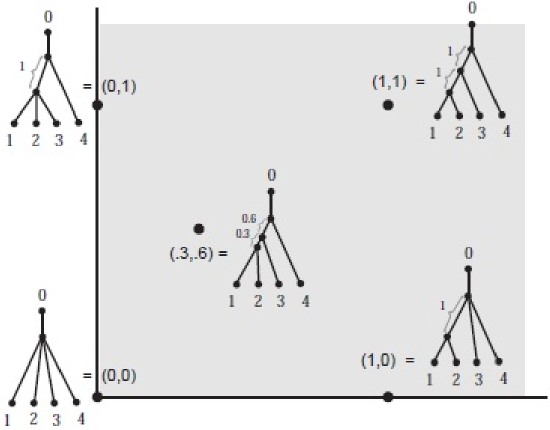


Figure 8: The 2-dimensional quadrant corresponding to a metric 4-tree

Trees with the same combinatorics form an orthant.

- A binary tree has the maximal possible number of interior edges($m-2$), and thus determines the largest possible dimensional orthant ($m-2$)
- The orthant corresponding to non-binary tree appears as a boundary face of at least 3 orthants corresponding to binary trees
- The origin of each orthant corresponds to the tree with no interior edges

We construct the space $\mathcal{T}_m$ by taking one $(m-2)$-dimensional orthant for each possible binary trees, and gluing them together along their common faces.

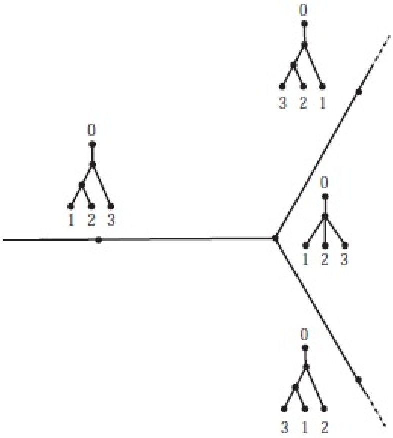


Figure 9: $\mathcal{T}_3$

We construct the space $\mathcal{T}_m$ by taking one $(m-2)$-dimensional orthant for each possible binary trees, and gluing them together along their common faces.

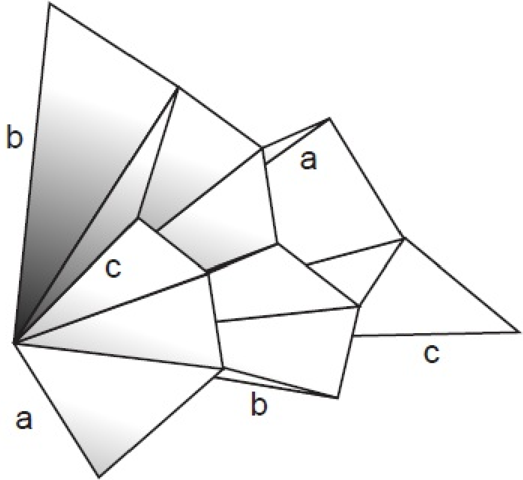


Figure 14: $\mathcal{T}_4$

Geodesic is the shortest path.

Definition

A geodesic from $x \in X$ to $y \in X$ is a map $c : [0, l] \subset \mathbb{R} \to X$ such that $c(0) = x, \; c(l) = y$ and $\forall t, t' \in [0, l], \; d(c(t), c(t')) = |t - t'|$
A geodesic segment from x to y is $[x, y] = c([0, l])$

CAT(0) is the generalization of non-positive curvature.

Definition

X is said to be $CAT(0)$ if the following is true:
$\forall a, b, c \in X$ with $d_1 = d(b, c)$, $d_2 = d(a, c)$ and $d_3 = d(a, b)$, form a "comparison triangle" in the Euclidean plane with vertices a', b' and c' with side length $d_1 = d(b', c')$, $d_2 = d(a,' c')$ and $d_3 = d(a', b')$. If $x \in [a, b]$, find $x' \in [a', b']$ with $d(a, x) = d(a', x')$.
Then $d(x, c) \leq d(x', c')$.

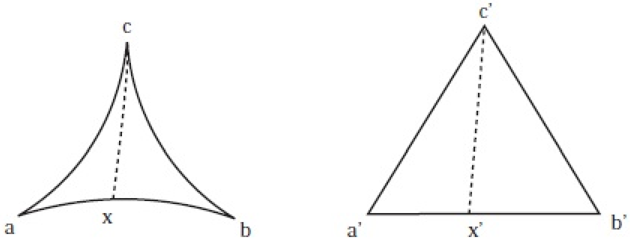


Figure 16: Comparison triangle

The space of trees has nonpositive curvature.

Lemma

([Billera et al., 2001] Lem 4.1) $\mathcal{T}_m$ is a CAT(0) space.

Nonpositive curvature space has unique geodesic.

- ([Bridson and Häfliger, 2011] Prop II.1.4) $\mathcal{T}_m$ being $CAT(0)$ implies that there exists unique geodesic segment connecting any two points of $\mathcal{T}_m$.

Fréchet mean is a generalization of average in metric spaces.

Definition

Given a probability measure μ on a tree space $\mathcal{T}_m$, its Fréchet mean T^* is

$$T^* = \underset{t \in \mathcal{T}_m}{\arg\min} \int_{\mathcal{T}_m} d(t, T)^2 d\mu(T).$$

Our goal is to characterize the limiting distribution of the sample Fréchet mean $\hat{T}$.

- For a collection of trees $T_1, \ldots, T_n \in \mathcal{T}_m$, the sample Fréchet mean $\hat{T}$ is the Fréchet mean on empirical measure, i.e.

$$\hat{T} = \arg\min_{t \in \mathcal{T}_m} \sum_{i=1}^{n} d(t, T_i)^2.$$

- We characterize the limiting distribution $\sqrt{n}(\hat{T} - T^*)$.

The log map is the generalisation of the inverse of the exponential map on a Riemannian manifold.

Definition

([Barden et al., 2014]) For a tree T^* in top dimensional orthant in $\mathcal{T}_m$, the log map $\log_{T^*} : \mathcal{T}_m \to \mathbb{R}^{m-2}$ at T^* takes the form

$$\log_{T^*}(T) = d(T^*, T) v_{T^*}(T),$$

where $v_{T^*}(T)$ is a unit vector at T^* along the geodesic from T^* to T.

- This is well-defined since $\mathcal{T}_m$ being CAT(0) implies that the geodesic is unique.

The modified log map adjusts the log map to originate from the base tree.

Definition

([Barden et al., 2014]) For a tree T^* in top dimensional orthant in $\mathcal{T}_m$, the modified log map $\Phi_{T^*} : \mathcal{T}_m \to \mathbb{R}^{m-2}$ at T^* takes the form

$$\Phi_{T^*}(T) = \log_{T^*}(T) + t^*,$$

for t^* the coordinates in $\mathbb{R}^{m-2}$ of T^*'s edge lengths.

The Fréchet mean of tree space is the average on the log space.

Lemma

([Barden et al., 2014], Lemma 3) Assume that the Fréchet mean T^* *of* μ *lies on a top dimensional orthant. Then* T^* *is characterized as*

$$\int_{\mathcal{T}_m} \Phi_{T^*}(T) d\mu(T) = T^*.$$

The limiting distribution of the sample Fréchet mean is Gaussian.

Theorem

([Barden et al., 2014], Theorem 2) Let μ be a probability measure on T_m with finite Fréchet function and with Fréchet mean T^* lying in a top-dimensional orthant. Assume that $\mu(\mathcal{D}) = 0$, where $\mathcal{D}$ is the set of trees with at least one internal branch of length zero. Suppose $\{T_i\}_{i\in\mathbb{N}}$ is a sequence of iid random variables in $\mathcal{T}_m$ with probability measure μ and denote by $\hat{T}_n$ the sample Fréchet mean of $T_1, \ldots, T_n$. Then

$$\sqrt{n}(\hat{T}_n - T^*) \rightsquigarrow \mathcal{N}(0, A^\top VA),$$

where V is the covariance matrix of the random variable $\Phi_{T^*}(T_1)$, and

$$A = (I - \mathbb{E}[M_{T^*}(T_1)])^{-1},$$

where $M_{T^*}(T)$ is the derivative of $\Phi_{T^*}(T)$ at T^*, with respect to T^*.

The limiting distribution of the sample Fréchet mean is Gaussian.

Proof.

$$
\begin{aligned}
\sqrt{n}(\hat{T}_n - T^*) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\Phi_{\hat{T}_n}(T_i) - T^*) \quad \text{(from Lemma)} \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\Phi_{T^*}(T_i) - T^*) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\Phi_{\hat{T}_n}(T_i) - \Phi_{T^*}(T_i)) \\
&\approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\Phi_{T^*}(T_i) - T^*) + \sqrt{n}(\hat{T}_n - T^*) \frac{1}{n} \sum_{i=1}^{n} M_{T_*}(T_i),
\end{aligned}
$$

hence

$$
\sqrt{n}(\hat{T}_n - T^*) \left(I - \frac{1}{n} \sum_{i=1}^{n} M_{T_*}(T_i) \right) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\Phi_{T^*}(T_i) - T^*).
$$

And then apply delta method and slutsky theorem. □

Reference

D. Barden, H. Le, and M. Owen. Limiting Behaviour of Fr\'echet Means in the Space of Phylogenetic Trees. *ArXiv e-prints*, September 2014.

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: http://dx.doi.org/10.1006/aama.2001.0759. URL //www.sciencedirect.com/science/article/pii/S0196885801907596.

M.R. Bridson and A. Häfliger. *Metric Spaces of Non-Positive Curvature*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2011. ISBN 9783540643241. URL https://books.google.com/books?id=3DjaqB08AwAC.

Thank you!