

GEODESICS IN THE SPACE OF TREES

KAREN VOGTMANN

1. INTRODUCTION

In a paper with Billera and Holmes published in 2001 we studied the space of finite metric trees with a fixed number of leaves, which can be interpreted as phylogenetic trees with positive branch lengths [1]. That paper included some information about how to find the geodesic path between two trees and the corresponding distance, but did not include a specific algorithm for doing so. The current notes address this task. They grew out of conversations with John Smillie, and were written shortly after the Billera-Holmes-Vogtmann paper appeared. They were never published, partly because I believe that a little more work (which I haven't done) could produce a substantially better algorithm. Meanwhile several people have requested these notes, so I decided to make them generally available.

Among the people who have asked for these notes, several have written computer programs to compute the distance between trees, including Aaron Staple (an undergraduate working with Susan Holmes at Stanford University), Megan Owen (a graduate student working with Lou Billera at Cornell) and Anne Kupczok (a PhD student at the Center for Integrative Bioinformatics Vienna).

2. TREES AND PARTITIONS

Removing a single edge from a tree separates that tree into two connected components, and therefore partitions the terminal vertices (called *leaves*) of the tree into two subsets. If the edge is not connected to a leaf (i.e. if the edge is *interior*), then each of these subsets has at least two elements. Furthermore, the combinatorial (i.e. simplicial) structure of the tree can be entirely reconstructed using just the information given by the partitions associated to all of its interior edges. These observations motivate the following definitions:

Definition. A partition of a finite set into two subsets, each with at least two elements, is called a *thick* partition. Two thick partitions $e = \{A_1|A_2\}$ and $f = \{B_1|B_2\}$ of a set X are *compatible* if some A_i is contained in some B_j or some A_i contains some B_j . If two partitions are not compatible, they are said to *cross*.

Any set E of pairwise compatible partitions determines a unique simplicial tree T_E .

Let E and F be sets of pairwise compatible thick partitions of the set $\{0, 1, \dots, n\}$, and let T_E and T_F be the corresponding rooted trees with n leaves, where the root is labeled by 0. Any such set of partitions can have at most $n - 2$ elements, and a set with $n - 2$ elements corresponds to a binary rooted tree. The root will not play a distinguished role in this discussion, and we think of it as simply another leaf.

3. THE CARRIER OF A GEODESIC

A set E of pairwise compatible thick partitions determines only the simplicial structure of a tree, not the metric structure. In tree space the edges of a tree have positive lengths, so that to specify a point in tree space you need both a set E of partitions and a positive real number $|e|$ associated to each partition $e = \{A|B\}$ in E . The exact path of the geodesic between two trees depends on both the simplicial structure *and* the metric structure of the tree. We will denote a metric tree with edges E by τ_E .

An orthant of tree space consists of all metric trees τ_E with underlying simplicial tree T_E , so we abuse notation and denote the orthant by T_E . The geodesic between two metric trees τ_E and τ_F in tree space passes sequentially through the interiors of orthants $T_E = T_0, T_1, \dots, T_{k-1}, T_k = T_F$. In [1] we showed that for any metric trees τ_E and τ_F , the minimal sequence of orthants containing the geodesic from τ_E to τ_F has the property that each T_i is obtained from T_{i-1} by removing a subset of E and adding a subset of F . We will call this sequence of orthants the *carrier* of the geodesic, and denote it $T(\tau_E, \tau_F)$.

The possibilities for the carrier of a geodesic γ are limited by the fact that the edge set E_i for each tree T_i must be a pairwise compatible subset of $E \cup F$.

4. TREES WITH A COMMON EDGE

Since edges of a tree correspond to partitions of the set $\{0, \dots, n\}$, it makes sense to talk about two different trees having “the same” edge. There is a proof in [1] that if E and F have an edge e in common, then every T_i in the carrier $T(\tau_E, \tau_F)$ contains e . In this case, we can simplify the problem of finding the geodesic as follows.

If e corresponds to the partition $\{A|B\}$, then $E = \{e\} \sqcup E(A) \sqcup E(B)$, where $E(A)$ consists of partitions $\{X|Y\}$ with $X \subset A$ or $Y \subset A$, and $E(B)$ is the partitions $\{X|Y\}$ with $X \subset B$ or $Y \subset B$.

Lemma 1. *Let $e = \{A|B\} \in E \cap F$. Then the carrier of the geodesic from τ_E to τ_F is the orthogonal product of the carrier for the geodesic between $\tau_{E(A)}$ and $\tau_{F(A)}$, the carrier for the geodesic between $\tau_{E(B)}$ and $\tau_{F(B)}$, and the orthant (ray) $T_{\{e\}}$:*

$$T(\tau_E, \tau_F) = T(\tau_{E(A)}, \tau_{F(A)}) \times T(\tau_{E(B)}, \tau_{F(B)}) \times T_{\{e\}}$$

Proof. In the tree $\tau_{E(B)}$, all of the leaves labeled by elements of A are attached at a single vertex. We can identify the subspace spanned by trees with all of the A 's attached at one vertex with the tree space on $|B|+1$ leaves; similarly, the subspace spanned by trees with all elements of B attached at one vertex can be identified with the tree space on $|A| + 1$ leaves. Each of these subspaces is isometrically embedded and independent of (orthogonal to) the other and to the one-dimensional orthant $T_{\{e\}}$. \square

Since the carrier is an orthogonal product, we can find the geodesic between τ_E and τ_F by separately finding the geodesics γ_A from $\tau_{E(A)}$ to $\tau_{F(A)}$, γ_B from $\tau_{E(B)}$ to $\tau_{F(B)}$, and taking $\gamma_e(t)$ to be the linear path from $|e|_E$ to $|e|_F$; then the geodesic $\gamma(t)$ is given by the formula

$$\gamma(t) = (\gamma_A(t), \gamma_B(t), \gamma_e(t)).$$

By applying this lemma repeatedly, we may reduce the problem of finding the geodesic from τ_E to τ_F to that of finding geodesics between trees with no edges in common. For the remainder of these notes we will therefore assume that $E \cap F = \emptyset$.

5. AN APPLICATION OF THE MARRIAGE LEMMA

We next reduce the problem of finding the geodesic between τ_E and τ_F to a problem in Euclidean space, using the following observation.

Lemma 2. *Let T_E and T_F be binary trees, with $E \cap F = \emptyset$. Then there is a bijection $b: E \rightarrow F$ such that e crosses $b(e)$ for all e .*

Proof. Any subset of $n - 1$ or more edges in $E \cup F$ must contain some incompatible pair, since a tree on n leaves can have at most $n - 2$ interior edges. Since all $n - 2$ edges of F are compatible, each edge of E must be incompatible with some edge in F . In fact, any set of k elements of E must be incompatible with k different edges of F ...otherwise we would could build a set with more than $n - 2$ compatible edges. We can now apply Philip Hall's marriage lemma to conclude that there is a way to match elements of E with elements of F in such a way that every matched pair is incompatible. (Kind of a cynical view of the Marriage Lemma). \square

Now let γ be the geodesic from τ_E to τ_F . Recall that the carrier of γ is contained in orthants corresponding to trees all of whose edges are in $E \cup F$.

Lemma 3. *The set of all orthants corresponding to trees with edges in $E \cup F$ can be embedded in \mathbb{R}^{n-2} as a union of orthants of \mathbb{R}^{n-2} , with the image of T_E in the all-negative orthant, and the image of T_F in the all-positive orthant. The image of γ is a geodesic in this image, with the intrinsic metric.*

Proof. Match each edge e_i of E with an incompatible edge f_i of F . For each i , send the ray determined by e_i in tree space to the negative ray of the x_i -axis and the ray determined by f_i to the positive x_i -axis. Then the image of e_i is orthogonal to the ray corresponding to any edge of $E \cup F$ with which it could possibly be compatible, so that every compatible set of k edges in $E \cup F$ determines a k -dimensional orthant of \mathbb{R}^{n-2} . Define the map to be an isometry on each such orthant. Since we know that the carrier of γ is contained in the union of these orthants, the last statement of the lemma follows. (You can decide whether a path is a geodesic based on local information, and all of the local information is in the image.) \square

We can now do all of our calculations in the Euclidean space \mathbb{R}^{n-2} , or rather in the set of orthants of \mathbb{R}^{n-2} which correspond to actual trees. We are trying to find the shortest path between a point in the all-negative orthant Q_- (the image of T_E) and one in the all-positive orthant Q_+ (the image of T_F), but we aren't allowed to enter orthants corresponding to incompatible subsets of $E \cup F$ (which we will call *illegal* orthants). One other piece of information we have is that the union of legal orthants is CAT(0) (i.e. non-positively curved); in particular, we know that there is a unique shortest path between a point of Q_- and a point of Q_+ .

6. FINDING THE GEODESIC

In this section we will fix a set of closed orthants of \mathbb{R}^{n-2} which includes both Q_- and Q_+ , and find the shortest path γ from a point $\mathbf{a} = -(a_1, \dots, a_{n-2})$ of Q_- to a point $\mathbf{b} = (b_1, \dots, b_{n-2})$ of Q_+ which stays in these orthants.

Lemma 4. *γ crosses each hyperplane $x_i = 0$ exactly once. If it intersects an orthant, the intersection is either a point or a straight line segment.*

Proof. If γ crossed a hyperplane twice, in points x and y , you could shorten γ by replacing the part of γ between x and y by the projection of γ onto the hyperplane.

If γ intersected a (closed) orthant in anything other than a point or single straight line segment, you could shorten it locally. \square

Now suppose that γ is parametrized by arc length, and let $L = \|\gamma\|$ be the length of γ . Let $x_i(t)$ be the i th coordinate, so $\gamma(t) = (x_1(t), \dots, x_{n-2}(t))$, with $x_i(0) < 0$ and $x_i(L) > 0$ for all i . The above lemma says that $\gamma(t)$ is piecewise linear in \mathbb{R}^{n-2} , and only changes direction possibly when some coordinate $x_i(t)$ changes from negative to positive ($\gamma(t)$ stays in an orthant until some coordinate changes from negative to positive. Some coordinates could change to zero, but at least one will have to become positive, otherwise we would have a geodesic traveling to a face, then continuing *in* the face).

Lemma 5. *The graph of $x_i(t)$ is straight except possibly where it intersects the t -axis. Its intersection with the t -axis is connected.*

Proof. The fact that the geodesic bends only when it changes orthants means that the graph of x_i is straight except at times s when some coordinate changes sign, becomes zero or becomes non-zero. At such a point, $x_j(s) = 0$ for some j . We have to see that, if $x_i(s) \neq 0$, then the graph of x_i is straight through $(s, x_i(s))$.

More generally, consider a path β path in \mathbb{R}^{n-2} consisting of a straight segment from a to p followed by a straight segment from p to b , and consider what happens to the length L of β when we vary the point p in a direction v , e.g. along $p(r) = p + rv$. We calculate

$$\frac{dL}{dr} = \left(\frac{p-a}{\|p-a\|} - \frac{p-b}{\|p-b\|} \right) \cdot v.$$

If $\beta = \beta(t)$ is parametrized by arc length, and $p = \beta(s)$, then the vector $\frac{p-a}{\|p-a\|}$ in the above expression is just $\frac{d\beta^-}{dt} = \frac{d\beta}{dt}|_{t \rightarrow s^-}$; similarly, $\frac{b-p}{\|b-p\|} = \frac{d\beta}{dt}|_{t \rightarrow s^+}$, and we can decide whether moving in the v direction decreases arc length by looking at the dot product

$$v \cdot \left(\frac{d\beta^-}{dt} - \frac{d\beta^+}{dt} \right)$$

We now return to our situation, with $x_i(s) \neq 0$ for some $i \neq j$. If γ is straight at s , then the graph of x_i is also straight at s , so suppose instead that γ bends at s . Since γ bends locally at isolated points, we can apply the above analysis to conclude that we could shorten γ at s by moving in any direction v with

$$v \cdot \left(\frac{d\gamma}{dt}|_{t \rightarrow s^-} - \frac{d\gamma}{dt}|_{t \rightarrow s^+} \right) < 0$$

In particular, if $\frac{dx_i}{dt}|_{t \rightarrow s^-} \neq \frac{dx_i}{dt}|_{t \rightarrow s^+}$, i.e. if the graph of x_i is not straight at $(s, x_i(s))$, then γ can be shortened by moving in the direction e_i or $-e_i$.

A move in the direction $\pm e_i$ changes only the i -th coordinates of points on the path. Since the paths are parametrized by arc length, the parametrization would also change, to $\gamma'(t) = (x'_1(t), \dots, x'_k(t))$, with $x'_k(s') = x_k(s)$ for all $k \neq i$ and for some s' close to s . Since $x_i(s) \neq 0$, a small enough move would not change the fact that $x'_i(s') \neq 0$, i.e. $\gamma(s)$ and $\gamma'(s')$ are in the same orthant, so that we can shorten the path without changing the sequence of orthants which γ passes through, and in particular without entering an “illegal” orthant of \mathbb{R}^{n-2} , i.e. one which does not correspond to a tree. Thus γ is not a geodesic.

The second statement of the lemma follows from the first. □

The proof above in fact shows that straightening the graph of any individual x_i shortens γ , as long as γ is parametrized by arc length. In particular, if the graph of x_i bends when it crosses the t -axis, then γ can be made shorter by sliding the endpoints of the intersection interval (or point) to the right or left to make the interval of intersection shorter (or the

graph straighter). However, we are constrained in how far we can slide endpoints by the fact that we must stay in the legal orthants, so γ cannot cross the hyperplanes $x_i = 0$ in arbitrary order, i.e. the graphs of the x_i cannot cross the t axis in arbitrary order.

For each i , let $[l_i, r_i]$ be the intersection of the graph of $x_i(t)$ with the t -axis. We call a point s on the t -axis a *turning point* of γ if $s = l_i$ or $s = r_i$ for some i .

Let $s_1 < s_2 < \dots < s_m$ be the turning points of γ , and for each $k = 1, \dots, m$ let L_k be the set of indices i such that $s_k = l_i$, and R_k the set of i such that $s_k = r_i$. Thus the L_k and the R_k each give a partition of $\{0, 1, \dots, n\}$. The fact that γ cannot enter illegal orthants gives constraints on the sets L_k and R_k . If the L_k and R_k correspond to a succession of legal orthants, we say the sets $\{L_k\}$ and $\{R_k\}$ form a *legal partition system*.

Notation. For $I = \{i_1, \dots, i_l\} \subset \{1, \dots, n\}$, let \mathbf{a}_I denote the vector $(a_{i_1}, \dots, a_{i_l})$, with entries the coordinates of \mathbf{a} in the indexing set I ; similarly, let $\mathbf{b}_I = (b_{i_1}, \dots, b_{i_l})$.

Lemma 6. *Suppose that $\{L_k\}$ and $\{R_k\}$ form a legal partition system of $\{0, 1, \dots, n\}$. If γ is the shortest path through the corresponding sequence of orthants, parametrized at constant speed from $t = 0$ to $t = 1$, then the turning points of γ are given by the formulas*

$$s_k = \frac{\|a_{L_k}\|}{\|a_{L_k}\| + \|b_{R_k}\|}$$

Proof. Any reparametrization of γ with constant speed is simply a rescaling of t , and does not affect the properties of the x_i established in the previous Lemma. (In fact, maybe I should state that lemma with regard to a constant speed parametrization, instead of a parametrization by arc length).

Fix k , and set $s = s_k$, $L = L_k$, $R = R_k$. Let λ_i be the slope of the graph of $x_i(t)$ to the left of s , and ρ_i the slope to the right of s . Since γ has constant speed, we have $\sum_{i=1}^n \lambda_i^2 = \sum_{i=1}^n \rho_i^2$. The only slopes which change at s are those of the x_i with $i \in L \cup R$, so in fact

$$\sum_{i \in L \cup R} \lambda_i^2 = \sum_{i \in L \cup R} \rho_i^2.$$

Now $\lambda_i = |a_i|/s$ if $i \in L$, and 0 otherwise; similarly, $\rho_i = |b_i|/(1-s)$ if $i \in R$ and 0 otherwise, so the above equation gives

$$\|a_L\|^2(1-s)^2 = \left(\sum_{i \in L} a_i^2\right)(1-s)^2 = \left(\sum_{i \in R} b_i^2\right)s^2 = \|b_R\|^2 s^2.$$

Solving this for s gives the result. □

Corollary 1. *The length of γ is equal to the square root of*

$$\sum_{i=1}^k (\|a_{L_i}\| + \|b_{R_i}\|)^2.$$

Proof. The length of γ is

$$\begin{aligned}
 L(\gamma) &= \int_0^1 \sqrt{\frac{dx_1^2}{dt} + \dots + \frac{dx_n^2}{dt}} \\
 &= \int_0^{s_1} \sqrt{\lambda_{L_1}^2 + \dots + \lambda_{L_k}^2} + \int_{s_1}^{s_2} \sqrt{\rho_{R_1}^2 + \lambda_{L_2}^2 + \dots + \lambda_{L_k}^2} + \dots \\
 &\quad + \int_0^{s_1} \sqrt{\rho_{R_1}^2 + \dots + \rho_{R_k}^2} \\
 &= \int_0^1 \sqrt{\lambda_{L_1}^2 + \dots + \lambda_{L_k}^2} \\
 &= \sqrt{\lambda_{L_1}^2 + \dots + \lambda_{L_k}^2},
 \end{aligned}$$

where $\lambda_L^2 = \sum_{i \in L} \lambda_i^2$ and $\rho_R^2 = \sum_{i \in R} \rho_i^2$. (This works since the path has constant speed.)
 Now

$$\lambda_L^2 = \sum_{i \in L} \lambda_i^2 = \sum_{i \in L} \frac{|a_i|^2}{\frac{\|a_L\|}{\|a_L\| + \|b_R\|}} = (\|a_L\| + \|b_R\|)^2,$$

which completes the proof of the lemma. \square

7. RECAP

We have shown: Geodesics in the sequence of orthants specified by $\{L_k\}$ and $\{R_k\}$ look like $\gamma(t) = (x_1(t), \dots, x_n(t))$ where, when parametrized at constant speed, the graph of $x_i(t)$ is a broken line segment, broken only at the points l_i and r_i , where it intersects the t -axis in the interval $[l_i, t_i]$.

If s_1, \dots, s_k are the turning points on the t -axis, then

$$s_j = \frac{\|a_{L_j}\|}{\|a_{L_j}\| + \|b_{R_j}\|},$$

and the total length of γ is the square root of

$$\sum_{i=1}^k (\|a_{L_i}\| + \|b_{R_i}\|)^2.$$

If we want to find the geodesic from \mathbf{a} to \mathbf{b} we proceed by finding all allowable decompositions of $\{1, \dots, n\}$ into $L_1 \sqcup \dots \sqcup L_k$, and $R_1 \sqcup \dots \sqcup R_k$, calculating the length of the geodesic through the corresponding orthants, and taking the one which is shortest. A decomposition is allowable if sequentially exchanging E_{L_j} for F_{R_j} results at each step in a compatible collection of edges, i.e. if $E - \bigcup_{j=1}^k E_{L_j} \cup \bigcup_{j=1}^k F_{R_j}$ is a compatible set of edges for each k .

8. REMARKS

This algorithm works, but it is exponential in the number of leaves of the tree. Once one decides on a sequence of orthants the distance computation is immediate from the formula in Corollary 1; the exponential nature of the algorithm comes from the fact that the number of possible legal sequences of orthants is exponential. My feeling is that it should be possible to use the fact that tree space is non-positively curved to reduce this to a polynomial-time algorithm.

Observations about the combinatorics and geometry of legal orthant-sequences allow you to eliminate many possibilities, but I don't see enough obvious reductions to get it down to a polynomial number.

REFERENCES

- [1] Louis J. Billera, Susan P. Holmes, and Karen Vogtmann. Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.*, 27(4):733–767, 2001.