Tropical Linear Spaces in Phylogenetics

Daniel Irving Bernstein

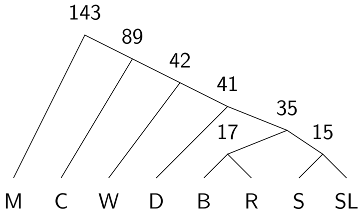
North Carolina State University
dibernst@ncsu.edu

https://arxiv.org/abs/1702.05141

May 27, 2016

Phylogenetics Motivation

The rooted tree below suggests an evolutionary history among eight species



	dog	bear	raccoon	weasel	seal	sea lion	cat	monkey
dog	0	41	41	42	41	41	89	143
bear	41	0	17	42	35	35	89	143
raccoon	41	17	0	42	35	35	89	143
weasel	42	42	42	0	42	42	89	143
seal	41	35	35	42	0	15	89	143
sea lion	41	35	35	42	15	0	89	143
cat	89	89	89	89	89	89	0	143
monkey	143	143	143	143	143	143	143	0

Phylogenetics Motivation

	dog	bear	raccoon	weasel	seal	sea lion	cat	monkey
dog	0	32	48	51	50	48	98	148
bear	32	0	26	34	29	33	84	136
raccoon	48	26	0	42	44	44	92	152
weasel	51	34	42	0	44	38	86	142
seal	50	29	44	44	0	24	89	142
sea lion	48	33	44	38	24	0	90	142
cat	98	84	92	86	89	90	0	148
monkey	148	136	152	142	142	142	148	0

Pairwise immunological distances between species[1]

Question

There is no way to display this dataset on a tree like the previous slide. What are the "closest" hypothetical datasets that can be?

[1] **Sarich 1969**.

Ultrametrics

Definition

A *dissimilarity map* on finite set X is a function $\mathbf{d} : X \times X \to \mathbb{R}$ such that

1. $\mathbf{d}(x, y) = \mathbf{d}(y, x)$ and
2. $\mathbf{d}(x, x) = 0$.

Equivalently, a dissimilarity map is a symmetric matrix with all zeros on the diagonal.

Definition

We say that $\mathbf{d}(\cdot, \cdot)$ is an *ultrametric* if for all $x, y, z \in X$ the maximum of $\mathbf{d}(x, y), \mathbf{d}(y, z)$ and $\mathbf{d}(x, z)$ is attained twice.

Proposition

A dissimilarity map can be expressed on a rooted tree if and only if it is an ultrametric. The leaf-labeled tree associated to an ultrametric is called its topology.

Motivating Question and Background

Question

Given a dissimilarity map **d**, which ultrametrics are nearest to **d** in the l^∞-norm? Do they all have the same topology?

In 2000, Chepoi and Fichet gave a polynomial-time algorithm for computing the coordinate-wise maximum ultrametric that is l^∞-nearest to a given dissimilarity map.

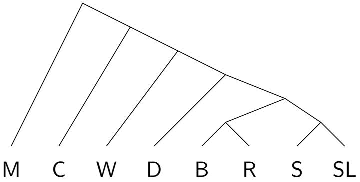
Theorem (Ardila and Klivans, 2006)

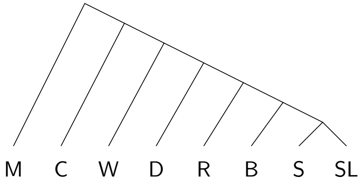
The set of ultrametrics on a finite set of size n is exactly the Bergman fan of the complete graph on n vertices.

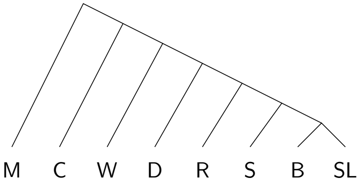
The above theorem gives a strong connection to tropical geometry which is one motivation for use of the l^∞-norm.

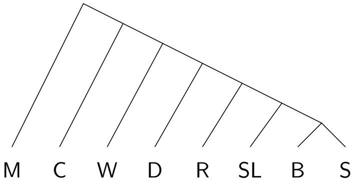
Multiple closest topologies

The ultrametrics closest in the l^∞-norm to Sarich's dataset are distance 9 away. Four different binary tree topologies are represented, shown below.









Question

What structure does the set of l^∞-nearest ultrametrics have?

Tropical Arithmetic

Definition

The *tropical semiring* is the extended real numbers $\mathbb{R} \cup \{-\infty\}$ where tropical addition is defined as

$$a \oplus b := \max\{a, b\}$$

and tropical multiplication is defined as

$$a \odot b := a + b.$$

Definition

The *tropical semi-module* is $(\mathbb{R} \cup \{-\infty\})^n$. If $x, y \in (\mathbb{R} \cup \{-\infty\})^n$, then $x \oplus y$ is the vector whose ith entry is $x_i \oplus y_i$. If $\alpha \in \mathbb{R} \cup \{-\infty\}$ then the ith entry of $\alpha \odot x$ is $\alpha \odot x_i$.

$$\begin{pmatrix}1\\2\end{pmatrix} \oplus 1 \odot \begin{pmatrix}-\infty\\2\end{pmatrix} = \begin{pmatrix}1\\2\end{pmatrix} \oplus \begin{pmatrix}-\infty\\3\end{pmatrix} = \begin{pmatrix}1\\3\end{pmatrix}$$

Tropical Convexity

In what follows, $v_1, \ldots, v_k \in (\mathbb{R} \cup \{-\infty\})^n$.

Definition

A tropical polytope is the tropical convex hull of a finite set of points. That is, a set of the form

$$\begin{aligned} \operatorname{tconv}(\{v_1, \ldots, v_k\}) := \{\lambda_1 \odot v_1 \oplus \cdots \oplus \lambda_k \odot v_k : \\ \lambda_1, \ldots, \lambda_k \in \mathbb{R} \cup \{-\infty\} \text{ and } \lambda_1 \oplus \cdots \oplus \lambda_k = 0\}. \end{aligned}$$

Definition

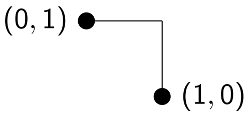
A tropical polyhedral cone is a set of the form

$$\begin{aligned} \operatorname{tcone}(\{v_1, \ldots, v_k\}) := \{\lambda_1 \odot v_1 \oplus \cdots \oplus \lambda_k \odot v_k : \\ \lambda_1, \ldots, \lambda_k \in \mathbb{R} \cup \{-\infty\}\}. \end{aligned}$$

Example

Consider the tropical polytope

$$\mathrm{tconv}\{(1,0),(0,1)\} \quad = $$



It contains the points

$$\begin{aligned}
(1,1) &= 0 \odot (1,0) \oplus 0 \odot (0,1) \\
(\frac{1}{2},1) &= (-\frac{1}{2} \odot (1,0)) \oplus (0 \odot (0,1)) \\
(1,\frac{1}{2}) &= (0 \odot (1,0)) \oplus (-\frac{1}{2} \odot (0,1))
\end{aligned}$$

(note $0 \oplus 0 = 0 \oplus -\frac{1}{2} = 0$).

Tropical polytopes and l^∞-nearest ultrametrics

Proposition (B. 2017)

The set of ultrametrics that are l^∞-nearest to a given dissimilarity map $\mathbf{d}$ *is a tropical polytope.*

Proposition (Develin-Sturmfels 2004)

Given a tropical polytope P, there exists a unique finite subset $V \subset P$ such that $P = \operatorname{tconv} V$. The elements of V are called the tropical vertices of P.

We give an algorithm that computes a superset of the tropical vertices of the set of ultrametrics l^∞-nearest to a given dissimilarity map.

Proposition (B. 2017)

All ultrametrics l^∞-nearest to $\mathbf{d}$ *have the same topology if and only if all the tropical vertices have the same topology.*

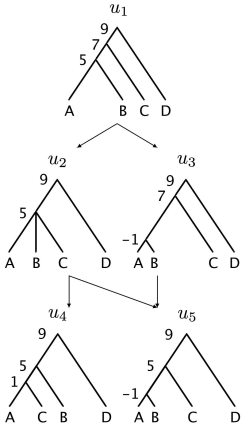
Algorithm

Algorithm idea:

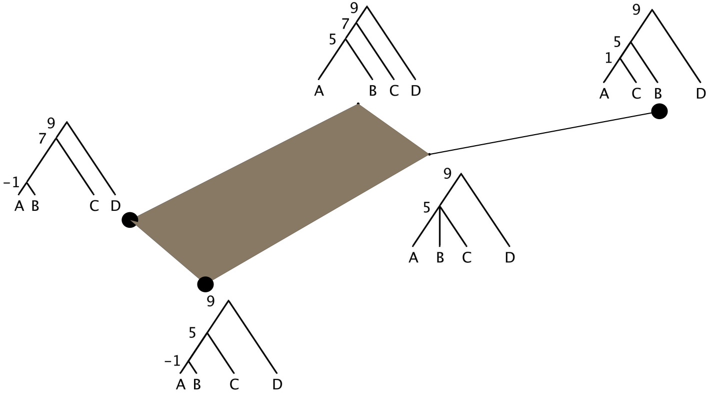
1. Compute maximal l^∞-nearest ultrametric using algorithm of Chepoi and Fichet
2. Slide internal nodes of the tree down until either creating a new polytomy, or sliding any further would increase l^∞ distance
3. Repeat
4. Return the ultrametrics such that at most one internal node can still be moved down

$$\mathbf{d} = \begin{array}{c c} & \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{array}{c} A \\ B \\ C \\ D \end{array} & \begin{pmatrix} 0 & 2 & 4 & 6 \\ 2 & 0 & 8 & 10 \\ 4 & 8 & 0 & 12 \\ 6 & 10 & 12 & 0 \end{pmatrix} \end{array}$$

d is distance 3 from nearest ultrametric



Tropical Polytope of Nearest Ultrametrics



Partial dissimilarity maps

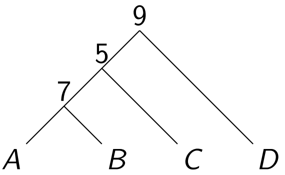
Question

Given a dissimilarity map where only some of the entries are known, is it possible to fill in the missing entries so that the result is an ultrametric?

$$
\begin{array}{cc}
 & \begin{array}{cccc} A & B & C & D \end{array} \\
\begin{array}{c} A \\ B \\ C \\ D \end{array} & \begin{pmatrix} 0 & 5 & \cdot & 9 \\ 5 & 0 & 7 & \cdot \\ \cdot & 7 & 0 & 9 \\ 9 & \cdot & 9 & 0 \end{pmatrix}
\end{array}
\qquad
\begin{array}{cc}
 & \begin{array}{cccc} A & B & C & D \end{array} \\
\begin{array}{c} A \\ B \\ C \\ D \end{array} & \begin{pmatrix} 0 & 5 & \cdot & 10 \\ 5 & 0 & 7 & \cdot \\ \cdot & 7 & 0 & 9 \\ 10 & \cdot & 9 & 0 \end{pmatrix}
\end{array}
$$

The partial dissimilarity map above on the left can be completed to an ultrametric while the one on the right cannot.

$$
\begin{array}{cc}
 & \begin{array}{cccc} A & B & C & D \end{array} \\
\begin{array}{c} A \\ B \\ C \\ D \end{array} & \begin{pmatrix} 0 & 5 & 7 & 9 \\ 5 & 0 & 7 & 9 \\ 7 & 7 & 0 & 9 \\ 9 & 9 & 9 & 0 \end{pmatrix}
\end{array}
$$



Bergman Fans

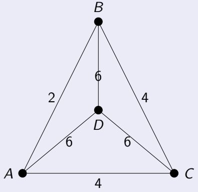
Definition

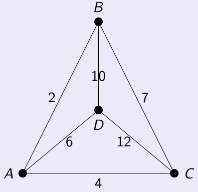
Let G be a graph with edge set E. The *Bergman fan of G* is

$$\tilde{\mathcal{B}}(G) := \{x \in \mathbb{R}^E : \text{if } C \text{ is a circuit of } G, \\ \text{then the maximum of } \{x_i : i \in C\} \text{ is attained twice}\}.$$

Example

Let K_4 be the complete graph on 4 vertices. The edge-labeled graph on the left represents an element of $\tilde{\mathcal{B}}(K_4)$ whereas the one on the right does not.





Bergman fans and (partial) ultrametrics

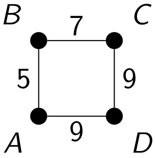
Theorem (Ardila and Klivans, 2006)

The set of ultrametrics on a finite set of size n is exactly the Bergman fan of the complete graph on n vertices.

Proposition (B. 2017)

Let $\mathbf{d}$ be a partial dissimilarity map supported on graph G. Then $\mathbf{d}$ can be completed to an ultrametric if and only if $\mathbf{d}$ lies on the Bergman fan of G.

$$\begin{array}{c c} & \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{array}{c} A \\ B \\ C \\ D \end{array} & \begin{pmatrix} 0 & 5 & \cdot & 9 \\ 5 & 0 & 7 & \cdot \\ \cdot & 7 & 0 & 9 \\ 9 & \cdot & 9 & 0 \end{pmatrix} \end{array}$$



$$\begin{array}{c c} & \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{array}{c} A \\ B \\ C \\ D \end{array} & \begin{pmatrix} 0 & 5 & 7 & 9 \\ 5 & 0 & 7 & 9 \\ 7 & 7 & 0 & 9 \\ 9 & 9 & 9 & 0 \end{pmatrix} \end{array}$$

Generalizing results on ultrametrics

Theorem (Ardila, 2004)

Bergman fans of graphs are tropical polyhedral cones. The Chipoi-Fichet algorithm for computing the coordinate-wise maximum l^∞-nearest ultrametric extends to Bergman Fans.

Theorem (B. 2017)

Let G be a graph with edge set E and let $x \in \mathbb{R}^E$. Then the subset of the Bergman fan $\tilde{\mathcal{B}}(G)$ of points that are l^∞-nearest to x is a tropical polytope. We have an algorithm for computing its tropical vertices.

Key idea for algorithm:

- Generalize notion of ultrametric topology so that elements of arbitrary Bergman fans can be assigned a topology
- All the pieces of the algorithm for ultrametrics now have a generalization

References

Federico Ardila.
Subdominant matroid ultrametrics.
Annals of Combinatorics, 8:379–389, 2004.

Federico Ardila and Caroline J. Klivans.
The Bergman complex of a matroid and phylogenetic trees.
Journal of Combinatorial Theory, Series B, 96(1):38 – 49,2006.

Daniel Irving Bernstein.
L-infinity optimization to Bergman fans of matroids with an application to phylogenetics.
`https://arxiv.org/abs/1702.05141`, **2017.**

Victor Chepoi and Bernard Fichet.
L-infinity optimization via subdominants.
Journal of Mathematical Psychology, 44, 600-616, 2000.

Mike Develin and Bernd Sturmfels.
Tropical convexity.
Documenta Mathematica, 9:1–27, 2004.

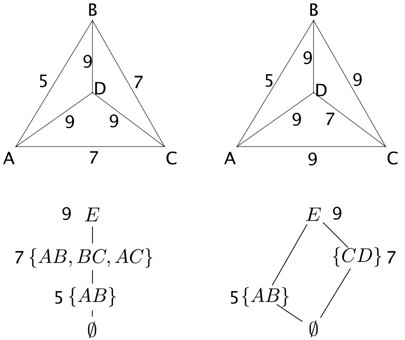
Vincent M Sarich.
Pinniped phylogeny.
Systematic Biology, 18(4):416–422, 1969.

Topology for Elements of Bergman Fans

Definition

A *hierarchy of connected flats* $\mathcal{F}$ of a matroid $\mathcal{M}$ is a collection of connected flats of $\mathcal{M}$ such that

1. $\emptyset \in \mathcal{F}$, and
2. If $F, G \in \mathcal{F}$ then $F \subseteq G$, $G \subseteq F$, or $F \cap G = \emptyset$



Proposition (B. 2017)

If $w \in \tilde{\mathcal{B}}(\mathcal{M})$ then there exists a unique hierarchy of connected flats $\mathcal{F}$ such that for each $F \in \mathcal{F}$

1. *w is constant on $F \setminus \bigcup_{\substack{G \in \mathcal{F} \\ G \subsetneq F}}$*
2. *If $G, F \in \mathcal{F}$ with $G \subsetneq F$, then $w_g < w_f$ for all $f \in F \setminus G$ and $g \in G$.*