Dissimilarity vectors of trees are contained in the tropical Grassmannian

Benjamin Iriarte Giraldo

Department of Mathematics, San Francisco State University
San Francisco, CA, USA

biriarte@sfsu.edu

Abstract

In this short writing, we prove that the set of m-dissimilarity vectors of phylogenetic n-trees is contained in the tropical Grassmannian $\mathcal{G}_{m,n}$, answering a question of Pachter and Speyer. We do this by proving an equivalent conjecture proposed by Cools.

1 Introduction.

This article essentially deals with the connection between phylogenetic trees and tropical geometry. That these two subjects are mathematically related can be traced back to Pachter and Speyer [7], Speyer and Sturmfels [9], and Ardila and Klivans [1]. The precise nature of this connection has been the matter of some recent papers by Bocci and Cools [2] and Cools [4]. In particular, a relation between m-dissimilarity vectors of phylogenetic n-trees with the tropical Grassmannians $\mathcal{G}_{m,n}$ has been noted.

Theorem 1.1 (Pachter and Sturmfels [8]). *The set of 2-dissimilarity vectors is equal to the tropical Grassmannian $\mathcal{G}_{2,n}$.*

This naturally raises the following question.

Question 1.2 (Pachter and Speyer [7], Problem 3). Does the space of m-dissimilarity vectors lie in $\mathcal{G}_{m,n}$ for $m \geqslant 3$?

The result in this article is of relevance in this direction and it is based on two papers of Cools [4] and Bocci and Cools [2], where the cases $m = 3$, $m = 4$ and $m = 5$ are handled. We answer Question 1.2 affirmatively for all m:

Theorem 1.3. *The set of m-dissimilarity vectors of phylogenetic n-trees is contained in the tropical Grassmannian $\mathcal{G}_{m,n}$.*

As we said, we prove Theorem 1.3 by proving an equivalent conjecture, Proposition 3.1 of this paper, or see Conjecture 4.4 of [4].

2   Definitions.

2.1   The Tropical Grassmannian.

Let $\mathbb{K} = \mathbb{C}\{\{t\}\}$ be the field of Puiseux series. Recall that this is the algebraically closed field of formal expressions

$$\omega = \sum_{k=p}^{\infty} c_k t^{k/q}$$

where $p \in \mathbb{Z}$, $c_p \neq 0$, $q \in \mathbb{Z}^+$ and $c_k \in \mathbb{C}$ for all $k \geqslant p$. It is the algebraic closure of the field of Laurent series over $\mathbb{C}$. The field comes equipped with a standard valuation val: $\mathbb{K} \mapsto \mathbb{Q} \cup \{\infty\}$ by which $\mathrm{val}(\omega) = p/q$. As a convention, $\mathrm{val}(0) = \infty$.

Now, let $x = (x_{ij})$ be an $m \times n$ matrix of indeterminates and let $\mathbb{K}[x]$ denote the polynomial ring over $\mathbb{K}$ generated by these indeterminates. Fix a second polynomial ring in $\binom{n}{m}$ indeterminates over the same field:

$$\mathbb{K}[p] = \mathbb{K}[p_{i_1, i_2, \ldots, i_m} : 1 \leqslant i_1 < i_2 < \cdots < i_m \leqslant n]$$

Let $\phi_{m,n} : \mathbb{K}[p] \mapsto \mathbb{K}[x]$ be the homomorphism of rings taking $p_{i_1, \ldots, i_m}$ to the maximal minor of x obtained from columns $i_1, \ldots, i_m$.

Definition 2.1. The *Plücker ideal* or ideal of *Plücker relations* is the homogeneous prime ideal $I_{m,n} = \ker(\phi_{m,n})$ which consists of the algebraic relations or syzygies among the $m \times m$ minors of any $m \times n$ matrix with entries in $\mathbb{K}$.

For $m \geqslant 3$, the Plücker ideal has a Gröbner basis consisting of quadrics; a comprehensive study of these ideals can be found in Chapter 14 of the book by Miller and Sturmfels [6] and in Sturmfels [10]. It is a polynomial ideal in $\mathbb{K}[p]$ and we can define its *tropical variety* in the usual way as we now recall. Let $a = \binom{n}{m}$ and $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. Consider

$$f = \sum c_\alpha p_{\sigma_1}^{\alpha_1} p_{\sigma_2}^{\alpha_2} \ldots p_{\sigma_a}^{\alpha_a} \in \mathbb{K}[p], \text{ where } \sigma_1, \ldots, \sigma_a \text{ are the } a \text{ } m\text{-subsets of } \{1, \ldots, n\}$$

The *tropicalization* of f is given by

$$\mathrm{trop}(f) = \min\{\mathrm{val}(c_\alpha) + \alpha_1 p_{\sigma_1} + \alpha_2 p_{\sigma_2} + \cdots + \alpha_a p_{\sigma_a}\}.$$

The *tropical hypersurface* $\mathcal{T}(f)$ of f is the set of points in $\overline{\mathbb{R}}^a$ where $\mathrm{trop}(f)$ attains its minimum twice or, equivalently, where $\mathrm{trop}(f)$ is not differentiable.

We are now ready to define tropical Grassmannians.

Definition 2.2. The tropical variety $\mathcal{T}(I_{m,n}) = \bigcap_{f \in I_{m,n}} \mathcal{T}(f)$ of the Plücker ideal $I_{m,n}$ is denoted by $\mathcal{G}_{m,n}$ and is called a *tropical Grassmannian*.

We have the following fundamental characterization of $\mathcal{G}_{m,n}$ which is a direct application of [9, Theorem 2.1].

Theorem 2.3. *The following subsets of $\overline{\mathbb{R}}^a$ coincide:*

- *The tropical Grassmannian $\mathcal{G}_{m,n}$.*

- *The closure of the set $\{(val(c_1), val(c_2), \ldots, val(c_a)) : (c_1, c_2, \ldots, c_a) \in V(I_{m,n}) \subseteq \mathbb{K}^a\}$*

2.2 Phylogenetic Trees.

We also treat phylogenetic trees in this paper.

Definition 2.4. A *phylogenetic n-tree* is a tree which has a labeling of its n leaves with the set $\{1, \ldots, n\}$ and such that each edge e has a positive real number $w(e)$ associated to it, which we call the *weight* of e.

There is also a crucial related family of trees which we now define:

Definition 2.5. An *ultrametric n-tree* is a binary rooted tree which has a labeling of its n leaves with $\{1, \ldots, n\}$ and such that

- each edge e has a nonnegative real number $w(e)$ associated to it, called the *weight* of e
- it is d-equidistant, for some $d > 0$, *i.e.* the sum of the edges in the path from the root to every leaf is precisely d
- the sum of the weights of all edges in the path connecting every two different leaves is positive.

Particularly, note that an ultrametric tree is binary and may have edges of weight 0. Now, let T be a phylogenetic n-tree. Define the vector $D(m, T)$ whose entries are the numbers d_σ, where σ is a subset of $\{1, 2, \ldots, n\}$ of size m and d_σ is the *total weight* of the smallest subtree of T which contains the leaves in σ. By the total weight of a tree, we mean the sum of the weights of all the edges in that tree.

Definition 2.6. The vector $D(m, T)$ is called the *m-dissimilarity vector of T*. The set of all m-dissimilarity vectors of phylogenetic trees with n leaves will be called the *space of m-dissimilarity vectors of n-trees*.

Definition 2.7. A metric space S with distance function $d : S \times S \mapsto \mathbb{R}_{\geq 0}$ is called an *ultrametric space* if the following inequality holds for all $x, y, z \in S$:

$$d(x, z) \leqslant \max\{d(x, y), d(y, z)\}$$

It is a well known fact that finite ultrametric spaces are realized by ultrametric trees, see for example [3, Lemma 11.1].

2.3 Column Reductions.

Let $n \geqslant 4$. Suppose we are given integers $1 \leqslant a, b \leqslant n$ with $a \neq b$ and let $c_{a,b}$ be the operator acting on Puiseux matrices for which, for any $n \times n$ matrix M, $c_{a,b}(M)$ is the matrix obtained from M by subtracting column b to column a. We know $c_{a,b}$ preserves the determinant, *i.e.* $\det(c_{a,b}(M)) = \det(M)$. For $l \geqslant 1$, let $(c_{a_l,b_l} \circ \cdots \circ c_{a_2,b_2} \circ c_{a_1,b_1})(M)$ be the matrix obtained from M by first subtracting column b_1 to column a_1, then subtracting column b_2 to column a_2, and so on up to subtracting column b_l to column a_l. Call this matrix a *column reduction of M* if the following conditions are met:

- $1 \leqslant a_1, \ldots, a_l, b_1, \ldots, b_l \leqslant n$

- the numbers $a_1, a_2, \ldots, a_l$ are pairwise different

- whenever $1 \leqslant k \leqslant l$, the number b_k is different from $a_1, \ldots, a_k$.

For simplicity, we will accept M as a column reduction of itself.

3 Main Result.

We are now ready to prove Theorem 1.3. Cools [4] reduced it to the following statement which we now prove.

Proposition 3.1 (Cools [4], Conjecture 4.4). *Assume $n \geqslant 4$. Let T be a d-equidistant ultrametric n-tree with root r and such that all its edges have rational weight.*

For each edge e of T, denote by $h(e)$ the well-defined sum of the weights of all the edges in the path from the top node of e to any leaf below e and let $a_1(e), \ldots, a_{n-2}(e)$ be generic complex numbers.

Let $x_i^{(j)} \in \mathbb{K}$ (with $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, n-2\}$) be the sum of the monomials $a_j(e)t^{-h(e)}$, where e runs over all edges between r and i. Then, the valuation of the determinant of

$$
M = \begin{pmatrix}
1 & 1 & \ldots & 1 \\
x_1^{(1)} & x_2^{(1)} & \ldots & x_n^{(1)} \\
(x_1^{(1)})^2 & (x_2^{(1)})^2 & \ldots & (x_n^{(1)})^2 \\
x_1^{(2)} & x_2^{(2)} & \ldots & x_n^{(2)} \\
\vdots & \vdots & \vdots & \vdots \\
x_1^{(n-2)} & x_2^{(n-2)} & \ldots & x_n^{(n-2)}
\end{pmatrix}
$$

is equal to $-D$, where D is the total weight of T.

In the course of the proof, we assume T is binary, which follows from the construction of Bocci and Cools [2]. Notice they start with a phylogenetic tree and then define an ultrametric associated with its 2-dissimilarity vector, therefore inducing an ultrametric tree. Here, T corresponds to certain subtrees of this induced ultrametric tree.

Proof. As T is binary, we know T has n leaves, $n - 2$ internal nodes of degree 3, 1 node (the root) of degree 2 and $2(n - 1)$ edges.

Let $\leqslant_T$ be the tree order of T with respect to r, *i.e.* the order on the set of nodes of T by which $v \leqslant_T w$ iff v lies in the path from r to w in T. Let $v_1, v_2, \ldots, v_{n-1}$ be the $n - 1$ internal nodes of T numbered in such way that if $v_i \leqslant_T v_j$, then $j \leqslant i$. We must have $v_{n-1} = r$.

Define an injective function $\alpha : v_i \mapsto a_i$ from the set of internal nodes to the leaves of T so that $v_i \leqslant_T a_i$ for all i with $1 \leqslant i \leqslant n - 1$. Now, for each of these values of i, let b_i be the unique leaf such that $b_i \neq a_j$ for all j with $1 \leqslant j \leqslant i$, and such that $v_i \leqslant_T b_i$.

If we calculate the column reduction $M^* = \left(c_{a_{n-1},b_{n-1}} \circ \cdots \circ c_{a_2,b_2} \circ c_{a_1,b_1}\right)(M)$ of M, then the valuation of the nonzero terms of the form $\prod_{i=1}^{n} M^*_{i,\sigma(i)}$ with $\sigma \in S_n$ in the sum

$$\det(M^*) = \sum_{\sigma \in S_n} \left(\mathrm{sgn}(\sigma) \prod_{i=1}^{n} M^*_{i,\sigma(i)} \right),$$

is precisely $-\left(\sum_{i=1}^{n-1} h(v_i) + d\right) = -D$. To see this notice for all i, $1 \leqslant i \leqslant n - 1$, we have

- $M^*_{1a_i} = 0$

- the valuation of $M^*_{3a_i}$ is $-d - h(v_i)$

- the valuation of $M^*_{ja_i}$ is $-h(v_i)$ if $j \neq 1$ and $j \neq 3$

- the only nonzero term in the first row of M^* is the 1 in column b_{n-1}

Because of our generic choice of coefficients, we can find some monomial term in the sum $\det(M^*)$ with valuation $-D$ which doesn't get cancelled, so we are done. $\square$

Example 3.2. Consider the 9-equidistant 10-tree of Figure 1 with total weight 35. The second row of the matrix M associated to this tree is the following vector with generic complex coefficients:

$$
\begin{aligned}
[&at^{-1} + ft^{-4} + pt^{-9} &&,bt^{-1} + ft^{-4} + pt^{-9} &&,ct^{-2} + gt^{-4} + pt^{-9} &&, \\
&dt^{-1} + ht^{-2} + gt^{-4} + pt^{-9} &&,et^{-1} + ht^{-2} + gt^{-4} + pt^{-9} &&,rt^{-1} + xt^{-3} + zt^{-4} + qt^{-9} &&, \\
&st^{-1} + xt^{-3} + zt^{-4} + qt^{-9} &&,ut^{-1} + yt^{-3} + zt^{-4} + qt^{-9} &&,vt^{-1} + yt^{-3} + zt^{-4} + qt^{-9} &&, \\
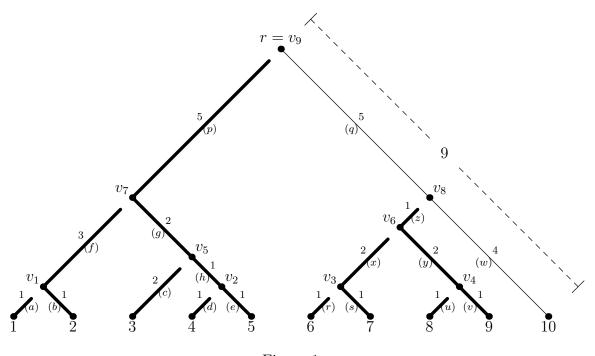&wt^{-4} + qt^{-9}]
\end{aligned}
$$

Figure 1:

A rooted 10-tree. The injective function
$$\alpha := \{(v_1, 1), (v_2, 4), (v_3, 6), (v_4, 8), (v_5, 3), (v_6, 7), (v_7, 2), (v_8, 9), (v_9, 5)\}$$
is depicted, as well as the equality $\sum_{i=1}^{9} h(v_i) = 35 - 9$.

Using the operator $(c_{5,10} \circ c_{9,10} \circ c_{2,5} \circ c_{7,9} \circ c_{3,5} \circ c_{8,9} \circ c_{6,7} \circ c_{4,5} \circ c_{1,2})$ suggested by the figure we obtain the column reduction M^* whose second row is the vector:

$$
\begin{aligned}
&[(a-b)t^{-1} &&,\quad (b-e)t^{-1} - ht^{-2} + (f-g)t^{-4} &&, \\
&\quad -et + (c-h)t^{-2} &&,\quad (d-e)t^{-1} &&, \\
&et^{-1} + ht^{-2} + (g-w)t^{-4} + (p-q)t^{-9} &&,\quad (r-s)t^{-1} &&, \\
&(s-v)t^{-1} + (x-y)t^{-3} &&,\quad (u-v)t^{-1} &&, \\
&vt^{-1} + yt^{-3} + (z-w)t^{-4} &&,\quad wt^{-4} + qt^{-9}]
\end{aligned}
$$

Also notice that $\sum_{i=1}^{9} h(v_i) = 35 - 9$.

We have shown that the m-dissimilarity vector of a phylogenetic tree T with n leaves gives a point in the tropical Grassmannian $\mathcal{G}_{m,n}$, and therefore gives rise to a tropical linear space. The combinatorial structure of those tropical linear spaces is the subject of an upcoming paper [5].

Acknowledgements.

References

[1] F. Ardila and C. Klivans, *The Bergman complex of a matroid and phylogenetic trees*, Journal of Combinatorial Theory, Series B, 96 (2006), 38-49.

[2] C. Bocci and F.Cools, *A tropical interpretation of m-dissimilarity maps*, Appl. Math. Comput. 212 (2009), 349–356.

[3] H-J. Böckenhauer and D. Bongartz, *Algorithmic aspects of bioinformatics*, Natural computing series, Springer-Verlag, Berlin Heidelberg, 2007.

[4] Filip Cools, *On the relation between weighted trees and tropical grassmannians*, J. Symb. Comput. 44 (2009), 1079–1086.

[5] B. Iriarte, *The tropical linear space of an m- dissimilarity vector*, in preparation.

[6] Ezra Miller and Bernd Sturmfels, *Combinatorial commutative algebra*, Graduate Texts in Mathematics, vol. 227, Springer-Verlag, New York, 2005.

[7] Lior Pachter and David Speyer, *Reconstructing trees from subtree weights*, Applied Mathematics Letters 17 (2004), 615–621.

[8] Lior Pachter and Bernd Sturmfels, *Algebraic statistics for computational biology*, Cambridge University Press, New York, 2005.

[9] David Speyer and Bernd Sturmfels, *The tropical Grassmannian*, Adv. Geom. **4** (2004), no. 3, 389–411.

[10] Bernd Sturmfels, *Algorithms in Invariant Theory*, Texts and Monographs in Symbolic Computation, Springer-Verlag, Vienna, 1993.