mathematics

MDPI

Article

Tropical Balls and Its Applications to K Nearest Neighbor over the Space of Phylogenetic Trees

Ruriko Yoshida

Naval Postgraduate School, 1411 Cunningham Road, Monterey, CA 93943-5219, USA; ryoshida@nps.edu; Tel.: +1-831-656-2973

Abstract: A tropical ball is a ball defined by the tropical metric over the tropical projective torus. In this paper we show several properties of tropical balls over the tropical projective torus and also over the space of phylogenetic trees with a given set of leaf labels. Then we discuss its application to the K nearest neighbors (KNN) algorithm, a supervised learning method used to classify a high-dimensional vector into given categories by looking at a ball centered at the vector, which contains K vectors in the space.

Keywords: classification; max-plus algebra; phylogenomics; ultrametrics

Citation: Yoshida, R. Tropical Balls and Its Applications to K Nearest Neighbor over the Space of Phylogenetic Trees. *Mathematics* **2021**, *9*, 779. https://doi.org/10.3390/math9070779

Academic Editor: Junseok Kim

Received: 9 March 2021
Accepted: 3 April 2021
Published: 5 April 2021

1. Introduction

A phylogenetic tree with a given set of leaf labels $[n] = \{1, \ldots, n\}$ is a weighted tree whose leaves have labels $[n]$ while their interior nodes do not have labels. In phylogenetics, leaves in a phylogenetic tree represent observable species $[n]$ in the current time and a tree represents an evolutionary relationship between these species in a given set $[n]$.

In order to study evolutionary histories of species or genes in terms of molecular clock, leaves in a phylogenetic tree represent a given set of observable species in the current time, internal nodes in the tree represent common ancestors and branch lengths in the tree present evolutionary time in a molecular clock. Since we assume that all species in the tree have the same most common ancestor (the root of a phylogenetic tree), a phylogenetic tree of a given set of species has the property that a distance from its root to each leaf is same for all leaves in the tree. We call such a rooted phylogenetic tree an *equidistant tree*. An example of an equidistant tree is shown in Figure 1. In phylogenetics and phylogenomics, we often use equidistant trees to analyze genome data since multispecies coalescent processes applied to analyze gene trees and species tree in genome data [1] assume that all gene trees are equidistant trees.

Phylogenomics is a field in which we apply tools from phylogenetics to problems in genomics. More specifically, phylogenomics extracts information from comparative study on entire genomes by constructing phylogenetic trees from each gene. In phylogenomics, researchers are interested in problems like predictions of gene function; evolutionary relationships between genes; and finding lateral gene transfers. For specific examples, genetic drift and gene flow in ancestral populations can cause topological differences between gene trees [1–4]. In the statistical point of view, these problems can be seen as finding outliers from a sample of phylogenetics with the same set of leaf labels reconstructed from genes. However, a space of phylogenetics trees (the set of all possible phylogenetic trees with a given set of leaf labels $[n]$) is not Euclidean. Thus, we cannot just simply apply statistical methods over a Euclidean space. Therefore, we have to consider a space of phylogenetic trees, which contains all possible phylogenetic trees of the same leaf set.

There are several ways to define a space of phylogenetic trees (also known as a *tree space*). These differences come from ways to define a phylogenetic tree and to *vectorize* each phylogenetic tree into a vector-representation (see [5] for details on differences between

different tree spaces). In this paper we focus on the space of phylogenetic trees as the set of all equidistant trees with a fixed set of labels for leaves.

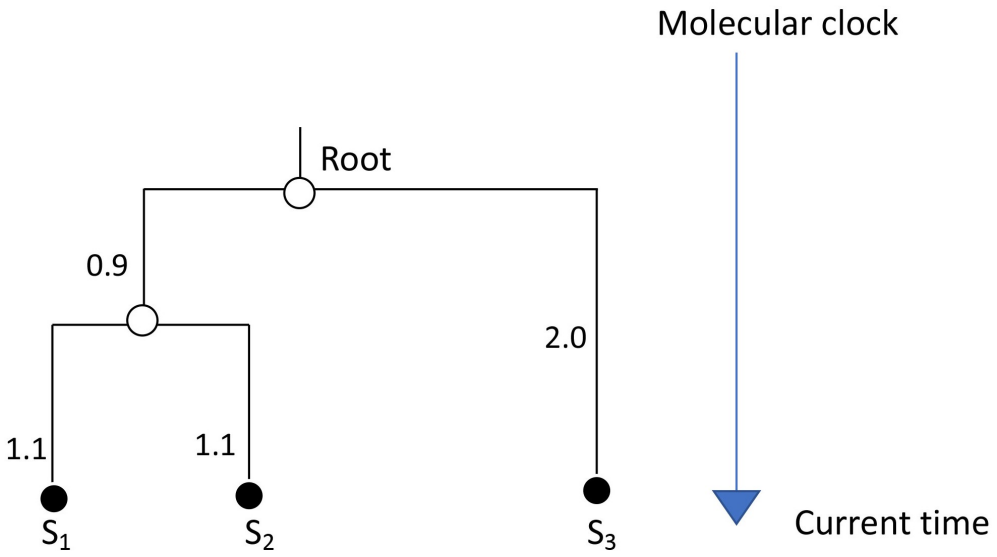

Figure 1. An equidistant tree with species S_1, S_2, S_3. Leaves in the tree represent observable species S_1, S_2, S_3 in the given set of labels and internal nodes in the tree represent their common ancestors. Filled black circles represent observable states and unfilled circles represent unobservable states. Number in each branch in the tree represent its branch length and the total branch lengths from the root to each leaf are same for all leaves.

In 2009, Speyer and Sturmfels showed a space of phylogenetic trees is a tropical Grassmanian [6], that is, a tropicalization of a linear space defined by a set of linear equations [7], which is a linear subspace of the tropical projective torus, with the max-plus algebra. Therefore it is natural to use tropical geometry to develop statistical methods to analyze data points over a space of phylogenetic trees. In this paper, we use the *tropical metric*. The tropical metric over a space of phylogenetic tree is well defined and [8,9] investigated its properties over the tree space. In 2020, Monod et al. in [10] introduced *tropical balls* over the space of phylogenetic trees with the tropical metric and defined probability measures using tropical balls. In [10], Monod et al. defined tropical balls with the tropical metric over the tropical projective torus. In this paper we show properties of tropical balls over the tropical projective torus with the tropical metrics as well as the space of phylogenetic trees. Then, we compare tropical balls with balls defined with L_2 norm and L_∞ norm.

The K Nearest Neighbors (KNN) algorithm is an instance-based method which classifies a vector in a high-dimensional vector space into finite categories [11]. The basic idea of the KNN algorithm is that all data points with the same category should be distributed near to each other. The KNN algorithm, first, considers the ball around the vector which you want to categorize. Next it increases the radius of the ball with a given metric until it contains K many points from the training set. Then we assign the category to the input vector which is the majority of data points from the training set in the ball. KNN algorithm is well-studied and it is applied to "Big Data" [12]. For example, [13] showed that the KNN algorithm is a special case of kernel density estimator.

The KNN algorithm has been applied to classify data points over a lower dimensional manifold (for example, [14]). However, since a space of phylogenetic trees is not a manifold over a Euclidean space and currently there has not been applied to classify data points over a space of phylogenetic trees. Therefore, in this paper, we consider *tropical KNN*, that is, KNN algorithm with the tropical metric defined over the tropical projective torus under the max-plus algebra to KNN and then we applied tropical KNN to classify data points over a space of phylogenetic trees.

The contributions of this paper include that

1. we show some properties of a tropical ball in the tropical projective torus;
2. we show some properties of a tropical ball in a space of equidistant trees with a given set of leaves $[n]$;
3. we compare tropical balls with balls defined with L_2 norm and L_∞ norm;
4. we define a tropical KNN algorithm; and
5. we applied tropical KNN algorithm to simulated data generated by a multispecies coalescent model.

2. Preliminaries

In this section, we set up some notation and definitions from phylogenetics and tropical geometry. For interested readers, see [15] for more details.

First we discuss some definitions from phylogenetics. A *dissimilarity map* w is a function $[n] \times [n] \to \mathbb{R}_{\geq 0}$ such that

$$w(i,j) = \begin{cases} w(j,i) \geq 0 & \text{if } i \neq j \\ 0 & \text{if } i = j. \end{cases}$$

If a dissimilarity map w satisfies a triangle inequality then w is called a *metric*. If there exists a phylogenetic tree T with the leaf labels $[n]$ such that the total branch length from a leaf $i \in [n]$ to a leaf $j \in [n]$ coincides with $w(i,j)$ for all $i, j \in [n]$, then we call w a *tree metric*. In fact if w is a tree metric of a phylogenetic tree T, then T has a unique tree metric w and T is unique for a tree metric w.

Definition 1. *If a metric w satisfies the following condition, such that*

$$\max\{w(i,j),\, w(i,k),\, w(j,k)\}$$

is achieved twice for every distinct $i, j, k \in [n]$, then we call a metric w an ultrametric.

To vectorize an equidistant tree, we can use an ultrametric w since there is a one-to-one mapping from a set of all possible equidistant trees with a leaf set $[n]$ to the space of all possible ultrametrics:

Theorem 1 ([10]). *A tree metric w of a phylogenetic tree T with a set of leaf labels $[n]$ is an ultrametric if and only if T is an equidistant tree with a set of leaf labels $[n]$.*

Therefore, in this paper, we define a space of equidistant trees with a leaf set $[n]$ as the *space of ultrametrics* notated as $\mathcal{U}_n$.

Now we shift our attentions to basics from tropical geometry. Throughout this paper we consider the max-plus tropical semiring $(\mathbb{R} \cup \{-\infty\}, \oplus, \odot)$.

Over this semiring, the tropical arithmetic operations of addition and multiplication are defined as the following:

$$x \oplus y := \max\{x, y\}, \quad x \odot y := x + y \quad \text{for any } x, y \in \mathbb{R} \cup \{-\infty\}.$$

Any semiring has to have the identity for addition and identity for multiplication. In this semiring, $-\infty$ is the identity for addition and 0 is the identity for multiplication.

Let $e := \binom{n}{2}$ and let $\mathbb{R}^e/\mathbb{R}\mathbf{1}$, where $\mathbf{1} := (1, 1, \ldots, 1)$, be the *tropical projective torus*. Note that $\mathcal{U}_n \subset \mathbb{R}^e/\mathbb{R}\mathbf{1}$. Scalar multiplication and vector addition can be defined as:

$$x \odot u = (x + u_1, x + u_2, \ldots, x + u_e)$$

$$x \odot u \oplus y \odot v = (\max\{x + u_1, y + v_1\}, \ldots, \max\{x + u_e, y + v_e\}),$$

where $x, y \in \mathbb{R} \cup \{-\infty\}$ and $u = (u_1, \ldots, u_e), v = (v_1, \ldots, v_e) \in (\mathbb{R} \cup -\{\infty\})^e$.

Through the paper, we use the *tropical metric*:

Definition 2. *For* $u = (u_1, \ldots, u_e)$, $v = (v_1, \ldots, v_e) \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$, *the* tropical distance d_{tr} *between* u *and* v *is defined as:*

$$d_{\mathrm{tr}}(u, v) = \max\{u_i - v_i : 1 \leq i \leq e\} - \min\{u_i - v_i : 1 \leq i \leq e\}. \quad (1)$$

Recall that from Theorem 1 we consider the *space of ultrametrics* with labels $[n]$ as a space of all equidistant trees with the label set $[n]$. Let $\mathcal{U}_n$ be the space of ultrametrics for equidistant trees with the leaf labels $[n]$. In fact we can write $\mathcal{U}_n$ as the tropicalization of the linear space generated by linear equations over the tropical projective torus $\mathbb{R}^e/\mathbb{R}\mathbf{1}$ where $e = \binom{n}{2}$.

Let $L_n \subseteq \mathbb{R}^e$ be the linear subspace of $\mathbb{R}^e$ defined by the linear equations:

$$w_{ij} - w_{ik} + w_{jk} = 0, \quad (2)$$

for $1 \leq i < j < k \leq n$. For the linear equations (2) spanning the linear space L_n, the max-plus tropicalization $Trop(L_n)$ of the linear space L_n is the tropical linear space with $w \in \mathbb{R}^e$ such that

$$\max\left\{w_{ij}, w_{ik}, w_{jk}\right\}$$

achieves at least twice for all $i, j, k \in [n]$. Note that this is exactly the three point condition defined in Definition 1.

Theorem 2 ([7], [Theorem 2.18])**.** $\mathcal{U}_n \subset \mathbb{R}^e/\mathbb{R}\mathbf{1}$ *is equal to* $Trop(L_n)$.

Therefore, by Theorem 2, $\mathcal{U}_n$ is a tropical linear space in $\mathbb{R}^e/\mathbb{R}\mathbf{1}$ where $e = \binom{n}{2}$. For example, if $n = 4$, the space of ultrametrics $\mathcal{U}_4$ is a union of 15 two-dimensional polyhedral cones over the tropical projective torus $\mathbb{R}^6/\mathbb{R}\mathbf{1}$.

3. Results

3.1. Properties of Tropical Balls over the Space of Ultrametrics

In this section we investigate the tropical ball in terms of the tropical metric in the space of ultrametrics $\mathcal{U}_n$. Here we fix the height of the equidistant trees associate with their ultrametrics in $\mathcal{U}_n$ as 1. All proofs for Lemmas and Theorems in this section are in Section 6. Recall $e := \binom{n}{2}$.

Lemma 1. *If* $u := (u_1, \ldots, u_e) \in \mathcal{U}_n$ *is from an ultrametric realizing an equidistant tree of height* 1, *then there is* $i \in \{1, \ldots, e\}$ *such that* $u_i = 2$. *In addition,*

$$\max(u) = \max_{1 \leq i \leq e}(u_i) = 2.$$

We will investigate a tropical ball centered at a point in $\mathcal{U}_n$ with a radius $r > 0$ defined via the tropical metric d_{tr}. We define a ball $B_x(r)$ at a point $x \in \mathcal{U}_n$ with a radius $r > 0$ under the tropical metric d_{tr} as the following:

$$B_x(r) = \{y \in \mathcal{U}_n : d_{\mathrm{tr}}(x, y) \leq r\}.$$

In the rest of this section, we show some properties on tropical balls in $\mathcal{U}_n$:

Proposition 1. *Suppose* $x \in \mathcal{U}_n$ *is the origin, that is, the ultrametric* $x = (0, 0, \ldots, 0) = (2, 2, \ldots, 2) \in \mathcal{U}_n$. *In terms of equidistant trees,* x *represents a star tree with its height* 1. *Then for* $0 < r \leq 2$,

$$B_x(r) = \{u \in \mathcal{U}_m : 2 - r \leq \min(u)\}.$$

Now we show that a tropical ball is convex in terms of the tropical metric. For Lemmas 2 and 3, and Theorem 3, let us consider the tropical projective torus $\mathbb{R}^e/\mathbb{R}\mathbf{1}$.

Lemma 2. *Suppose $x, z \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$ and $a \in \mathbb{R}$. Then*

$$d_{\mathrm{tr}}(a \odot x, z) = d_{\mathrm{tr}}(x, z).$$

Lemma 3. *Suppose $x, y, z \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$ and $a \in \mathbb{R}$. Then*

$$d_{\mathrm{tr}}(x \oplus y, z) \leq \max\{d_{\mathrm{tr}}(x, z), d_{\mathrm{tr}}(x, z)\}.$$

Theorem 3. *A tropical ball $B_x(r)$ in the tropical projective torus $\mathbb{R}^e/\mathbb{R}\mathbf{1}$ is convex under the tropical metric d_{tr}.*

Since $\mathcal{U}_m$ is a tropical linear space over the tropical projective torus $\mathbb{R}^e/\mathbb{R}\mathbf{1}$ by Theorem 2 and since $B_x(r)$ is convex, now we have the following theorem by Theorem 3.

Theorem 4. *A tropical ball $B_x(r)$ in $\mathcal{U}_m$ is convex in terms of the tropical metric d_{tr}.*

Until this moment in this section, we consider ultrametrics in $\mathcal{U}_n \subset \mathbb{R}^e/\mathbb{R}\mathbf{1}$. Even with a natural bijection between ultrametrics in $\mathcal{U}_n$ and the space of equidistant trees described in Theorem 1, one might want to describe a tropical balls in terms of equidistant trees. In general we can describe a tropical ball in $\mathcal{U}_n$ in terms of equidistant trees using the following theorem:

Theorem 5. *Let $\mathcal{T}_n(h)$ be the set of all equidistant trees with their height $h > 0$ and with the leaf set $[m] = \{1, \ldots, n\}$. Also let $d_T(i, j)$ be a pairwise distance between leaves $i, j \in \{1, \ldots, n\}$ in an equidistant tree $T \in \mathcal{T}_n(h)$. Suppose $T_1, T_2 \in \mathcal{T}_n(h)$. Then let $M(T_1, T_2) = \max\{d_{T_1}(i, j) - d_{T_2}(i, j) : i, j \in [n]\}$. For an equidistant tree $T_x \in \mathcal{T}_n(h)$ and for $0 < r \leq 4h$,*

$$B_{T_x}(r) = \{T_u \in \mathcal{T}_n(h) : M(T_x, T_u) + M(T_u, T_x) \leq r\}.$$

3.2. Examples in $\mathcal{U}_4$

In this section, we visualize tropical balls in the space of ultrametrics and in order to visualize, we consider $n = 4$. To visualize tropical balls in $\mathcal{U}_4$, we map all coordinates in $\mathcal{U}_4$ to the Billera-Holmes-Vogtmann (BHV) treespace [16] using the one-to-one mapping described in [17].

Example 1. *First, we consider the tree shown in the left side of Figure 2. This is an example so that the center of a tropical ball and entire tropical ball are inside of an orthant for a tree topology in Figure 2.*

Here we have the height $h = 1$, so

$$0 \leq a,\ b,\ a + b \leq 1.$$

The corresponding ultrametric is

$$u = (2(1 - a - b),\ 2(1 - b),\ 2,\ 2(1 - b),\ 2,\ 2).$$

The center of this tropical ball in this example is $a = b = 1/4$. The corresponding ultrametric is

$$x_0 = (1, 3/2, 2, 3/2, 2, 2).$$

Therefore, we have

$$u - x_0 = (1 - 2a - 2b, 1/2 - 2b, 0, 1/2 - 2b, 0, 0).$$

The tropical ball in $\mathcal{U}_4$ mapped in the BHV tree space is shown in the right side of Figure 2.

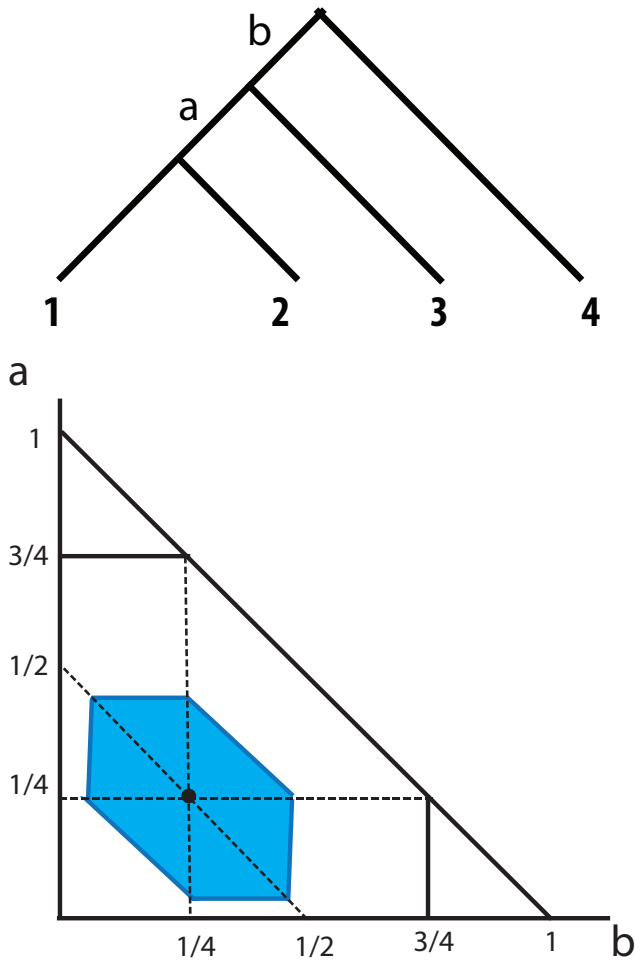

Figure 2. The first example for visualizing a tropical ball. LEFT: The tree corresponding to the center of the tropical ball. RIGHT: The tropical ball centered around the ultrametric corresponding to the equidistant tree in $\mathcal{U}_4$.

Example 2. *For the second example, we consider the tree shown in the left side of Figure 3. This is also a tropical ball contained inside of an orthant for the tree topology shown in Figure 3.*

Here we have the height $h = 1$, so

$$0 \leq a, b \leq 1.$$

The corresponding ultrametric is

$$u = (2(1-a), 2, 2, 2, 2, 2(1-b)).$$

The center of this tropical ball in this example is $a = b = 1/2$. The corresponding ultrametric is

$$x_0 = (1, 2, 2, 2, 2, 1).$$

Therefore, we have

$$u - x_0 = (2(1-a) - 1, 0, 0, 0, 0, 2(1-b) - 1).$$

The tropical ball around the ultrametric corresponding the equidistant tree is shown in the right side of Figure 3.

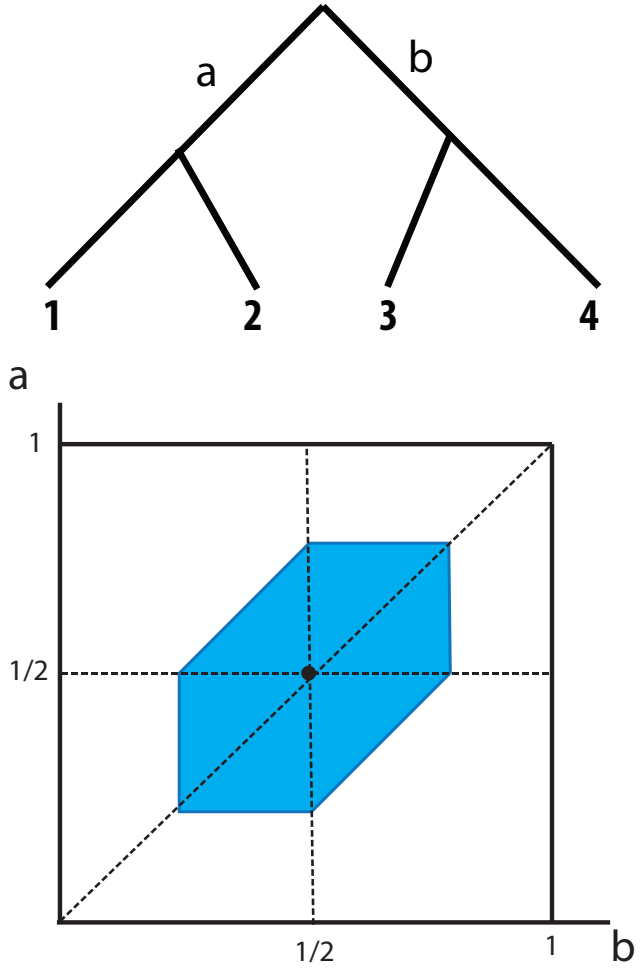

Figure 3. The second example for visualizing a tropical ball. LEFT: The tree corresponding to the center of the tropical ball. RIGHT: The tropical ball centered round the ultrametric corresponding to the equidistant tree in $\mathcal{U}_4$.

Example 3. *The third example we consider is given in the left side of Figure 4. This example shows a case that the tropical ball crosses between two orthants for two tree topologies shown in Figure 4.*

The center tree of the tropical ball in this example is the tree where $b = 0$, $c = 0$. This is the tree on the boundary of two orthants in the space of phylogenetic trees in this case. Here we have the height $h = 1$, so

$$0 \leq a_1,\, a_2,\, b,\, c,\, a_2 + c \leq 1.$$

The corresponding ultrametrics are

$$\begin{array}{rcl} u & = & (2(1-a_1),\, 2,\, 2,\, 2,\, 2,\, 2(1-b)) \\ v & = & (2(1-a-b),\, 2(1-b),\, 2,\, 2(1-b),\, 2,\, 2). \end{array}$$

The center of the tropical ball in this example is $a_1 = a_2 = 1/2, b = c = 0$. The corresponding ultrametric is

$$x_0 = (1, 2, 2, 2, 2, 2).$$

Therefore, we have

$$\begin{array}{rcl} u - x_0 & = & (2(1-a_1) - 1, 0, 0, 0, 0, -2b) \\ v - x_0 & = & (1 - 2a_2 - 2c, -2c, 0, -2c, 0, 0). \end{array}$$

The tropical ball centered around the ultrametric corresponding to the equidistant tree where $b = 0$ and $c = 0$ is shown in the right side of Figure 4.

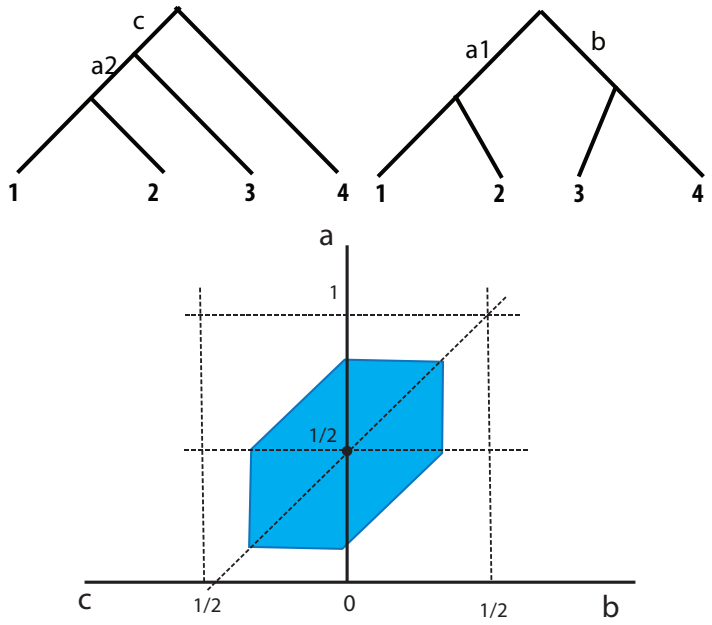

Figure 4. The third example for visualizing a tropical ball. LEFT: The tree corresponding to the center of the tropical ball is the tree where $b = 0$ and $c = 0$ in the picture. In this case the center of the tropical ball is on the boundary between two orthants in the tree space. RIGHT: The tropical ball centered around the ultrametric corresponding to the equidistant tree in $\mathcal{U}_4$.

3.3. Approximation of a Tropical Ball

In this subsection we discuss how a tropical ball with the radius $r > 0$ relates with a ball under the l_2 and l_∞ metrics. Suppose we have $u, v \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$. Let $d_2(u, v)$ be the l_2 norm metric and $d_\infty(u, v)$ be the l_∞ norm between x and y.

Suppose we can assume that

$$(u - v) > 0 \text{ and } \min(u - v) = 0$$

for $u, v \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$. Then we have the following proposition:

Proposition 2. *We assume that* $(u - v) \geq 0$ *over* $\mathbb{R}^e/\mathbb{R}\mathbf{1}$ *and* $\min(u - v) = 0$. *Then we have*

$$d_{\mathrm{tr}}(u, v) = d_\infty(u, v).$$

Proof.

$$\begin{aligned} d_\infty(u, v) &= \max(|u - v|) \\ &= \{\max(u - v), \max(v - u)\} \\ &= \max(u - v) \text{ since } (u - v) \geq 0 \\ &= \max(u - v) - \min(u - v) \text{ since } \min(u - v) = 0 \\ &= d_{\mathrm{tr}}(u, v). \end{aligned}$$

□

In this case the tropical ball coincides with the ball defined with the l_∞ norm. However, if we are working on the space of ultrametrics, we have the constraints that all points have to be ultrametrics and no longer we can assume that $(u - v) > 0$ and $\min(u - v) = 0$ in $\mathcal{U}_n$. Therefore for more general cases, we have the following bounds:

Proposition 3. *Suppose we have* $u, v \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$. *Then we have*

$$d_{\mathrm{tr}}(u, v) \leq 2 \cdot d_\infty(u, v) \leq 2 \cdot d_2(u, v).$$

Proof. The second inequality is trivial. Thus we want to prove that

$$d_{\mathrm{tr}}(u,v) \leq 2 \cdot d_{\infty}(u,v).$$

We can define $d_{\mathrm{tr}}(u,v)$ as

$$\begin{array}{rcl} d_{\infty}(u,v) & = & \max(|u-v|) \\ & = & \max\{\max(u-v), \max(v-u)\} \\ & = & \max\{\max(u-v), -\min(u-v)\}. \end{array}$$

and

$$d_{\mathrm{tr}}(u,v) = \max(u-v) - \min(u-v).$$

Therefore

$$\begin{array}{rcl} d_{\mathrm{tr}}(u,v) & = & \max(u-v) - \min(u-v) \\ & \leq & \max\{\max(u-v), -\min(u-v)\} + \max\{\max(u-v), -\min(u-v)\} \\ & = & 2 \cdot d_{\infty}(u,v). \end{array}$$

□

Using Proposition 3 we have the following theorem:

Theorem 6. *Let $B_r^2(x)$ be a ball around a point $x \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$ with the radius $r > 0$ with the l_2 metric d_2 and let $B_r^{\infty}(x)$ be a ball around a point $x \in R^e$ with the radius $r > 0$ with the l_2 metric d_{∞}. Then we have*

$$B_r(x) \subset B_{2r}^{\infty}(x) \subset B_{2r}^2(x).$$

Using Theorem 6 we can approximate a tropical KNN algorithm using the classical KNN algorithm with d_2 metric. We show some computational experiments using the multi-species coalescent model in the following section.

3.4. Computational Results

By Theorem 6 we can approximate a tropical ball using d_2 metric. Since the KNN algorithm uses the notion of balls to classify each data point in a test set, we can approximate a tropical KNN algorithm using the classical KNN algorithm. In this section we compare the tropical KNN algorithm and the classical KNN algorithm using d_2 with simulated data sets generated under the multi-species coalescent model via the software `Mesquite` [18]. All computations are conducted in Apple Notebook MacBook Pro 2019 with 2.4 GHz 8-Core Intel Core i9 and 64 GB 2667 MHz DDR4. The `R` code used for this simulation study can be found at polytopes.net/tropical_KNN.tar.

First we review the KNN algorithm. The KNN algorithm is a classification model for a data set where the response variable is categorical with finite levels and explanatory variables are all numerical. Suppose we have a data set $\{(c_1, x_1), \ldots (c_m, x_m)\}$ where $c_i \in \{1, \ldots, C\}$ with $C > 0$ positive integer and $x_i \in \mathbb{R}^e$. The outline of the Algorithm 1 is following:

Algorithm 1: KNN Algorithm.

- Input: A data point $y \in R^e$ from a test set, a training set $\{x_1, \ldots x_m\}$, and a metric d. Positive integer $k > 1$.
- Output: A class for y.
- Algorithm:

  for $i = 1, \ldots, m$ **do**
  Compute $d(y, x_i)$
  end for

  Order $d(y, x_1), \ldots, d(y, x_m)$ from the smallest to the largest. Suppose $d(y, x_{i_1}), \ldots, d(y, x_{i_k})$ be the first k smallest distances.

  Consider categories of $x_{i_1}, \ldots x_{i_k}$, that is, $c_{i_1}, \ldots, c_{i_k}$ and assign the class, which is the biggest frequency among $c_{i_1}, \ldots, c_{i_k}$, to y.

In terms of balls, basically we can think of the KNN algorithm as a method to find a ball around y, which contains k many points from the training set. From this view we can think that the KNN algorithm with the tropical metric d_{tr} can be approximated using the KNN algorithm with the l_2 metric d_2. For this simulation we implemented the tropical KNN, the KNN algorithm with the tropical metric in `R` and we use the KNN algorithm implemented in the "class" package in `R` [19].

For simulated data sets, we use the software `Mesquite`, available at http://mesquiteproject.org [18], to generate gene trees under the multispecies coalescent model. In this model, we have two parameters, the effective population size N_e and species depth SD. In this simulation study, we set $N_e = 100,000$ and varied

$$c = \frac{SD}{N_e}.$$

For this simulation, we generated species trees under the Yule model and gene trees given the species tree are generated by the multispecies coalescent model via `Mesquite`.

In computational experiments, we have varied $c = 0.25, 0.5, 1, 2, 5, 10$. In addition, we set $n = 10$.

With the knn() function from the "class" package, it took 12.7 s to finish the computations while the tropical metric took several hours since we can speed up the KNN algorithm for the l_2 metric. The results are shown in Figure 5.

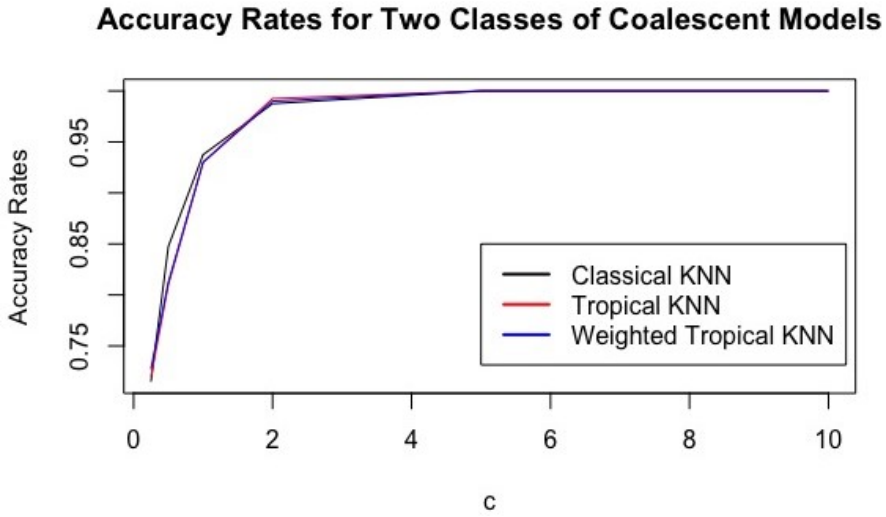

Figure 5. Accuracy Rates for the classical KNN, tropical KNN, and weighted tropical KNN on simulated coalescent models.

For the weighted tropical KNN, we assigned weight when we compute the majority votes such that

$$\frac{1}{(d_{\mathrm{tr}}(y, x_i))^2}.$$

4. Conclusions

In this paper, we discussed a tropical ball over the tropical projective torus and the space of ultrametrics. Then we discussed approximation of tropical balls using the l_2 and l_∞ metrics.

Then we discussed applications of the KNN algorithm and we showed by simulations, using the multi-species coalescent model, that approximation of tropical balls with the l_2 metric works well. In addition, we can consider the ensemble model, that is, taking the average of all three methods we used in the simulation study. This will increase the accuracy of the simulation study.

Since the tropical metric considers maximum and minimum of elements in a vector, this metric might be very sensitive to outliers. Therefore, we recommend to conduct analysis on outliers before applying a statistical method with the tropical metric.

5. Discussion

We focused on tropical balls with the tropical metric under the max-plus algebra in this paper. However, we still have many problems to solve. For example, we do not know how to compute a tropical ball over the tropical projective torus and the space of ultrametrics in general even though we can compute some small examples by hand. Explicitly,

Problem 1. *Develop an algorithm to compute a tropical ball around a point $x \in \mathbb{R}^e/\mathbb{R}$ with the radius $r > 0$, $B_r(x) \in \mathbb{R}^e/\mathbb{R}$.*

Problem 2. *Develop an algorithm to compute a tropical ball around a point $x \in \mathcal{U}_n$ with the radius $r > 0$, $B_r(x) \in \mathcal{U}_n$.*

In addition, we might want to consider a distribution-based clustering method, such as Density-based spatial clustering of applications with noise (DBSCAN) with the tropical metric for clustering gene trees. Similarly to the results on the KNN, with initial experiments, DBSCAN works fairly well on the data sets generated under the coalescent model shown in Figure 6.

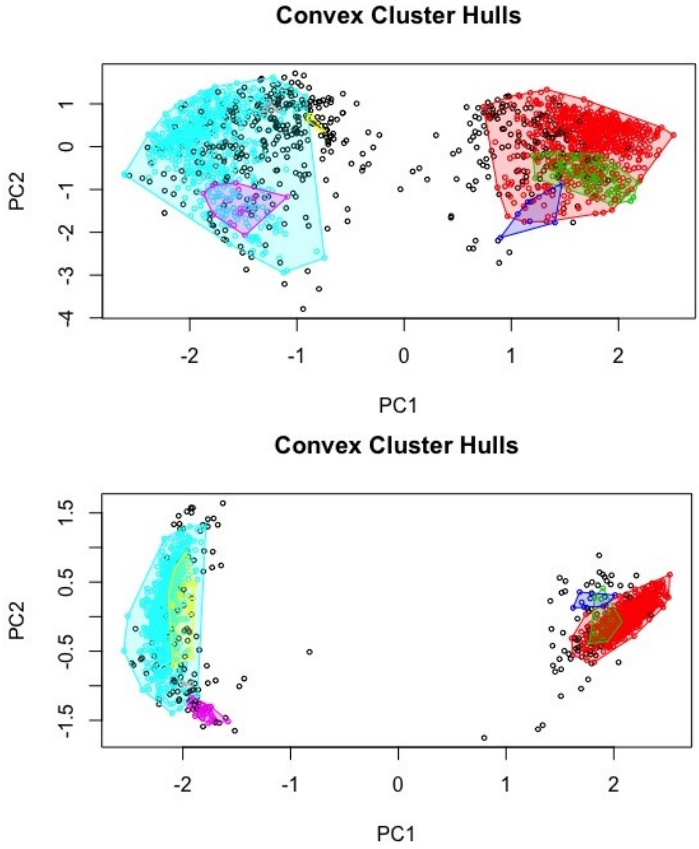

Figure 6. DBSCAN results for $c = 5$ (**Top**) and $c = 10$ (**Bottom**). The minpt = 5 for both cases, and esp = 1 for $c = 5$ and eps = 0.5 for $c = 10$.

6. Materials and Methods

In this section, we show proofs for propositions and theorems in Section 3.

Proof for Lemma 1. Since the root in the equidistant tree has degree of at least two, at least one pairwise distance in u is realized by a path through the root. In addition since the height of an equidistant tree is 1, therefor, the maximum of the pairwise distance from any leaf i to any leaf j is 2. □

Proof for Theorem 5. Let $x \in \mathcal{U}_m$ be an ultrametric associated with the equidistant tree $T_x \in \mathcal{T}_m(h)$ and let $u \in \mathcal{U}_m$ be an ultrametric associate with the equidistant tree $T_u \in \mathcal{T}_m(h)$. Then we have

$$d_{\mathrm{tr}}(x, u) = \max(x - u) - \min(x - u) \leq r.$$

Thus,

$$M(T_x, T_u) + M(T_u, T_x) \leq r.$$

□

Proof for Lemma 2. Let $x = (x_1, \ldots, x_e)$, $z = (z_1, \ldots, z_e)$. Then we have

$$a \odot x = (a + x_1, a + x_2, \ldots, a + x_e).$$

Also we have

$$\begin{aligned} \max(a \odot x - z) &= a + \max_i(x_i - z_i), \\ \min(a \odot x - z) &= a + \min_i(x_i - z_i). \end{aligned}$$

Thus we have

$$\begin{aligned} d_{\mathrm{tr}}(a \odot x, z) &= \max(a \odot x - z) - \min(a \odot x - z) \\ &= a + \max_i(x_i - z_i) - (a + \min_i(x_i - z_i)) \\ &= \max_i(x_i - z_i) - \min_i(x_i - z_i) \\ &= d_{\mathrm{tr}}(x, z). \end{aligned}$$

□

Proof for Lemma 3. Let $x = (x_1, \ldots, x_e)$, $y = (y_1, \ldots, y_e)$, $z = (z_1, \ldots, z_e)$. We have

$$\begin{aligned} d_{\mathrm{tr}}(x \oplus y, z) &= \max_i\{\max(x_i, y_i) - z_i\} - \min_i\{\max(x_i, y_i) - z_i\} \\ &= \max_i\{\max(x_i - z_i, y_i - z_i)\} - \min_i\{\max(x_i - z_i, y_i - z_i)\}, \end{aligned}$$

and

$$\max\{d_{\mathrm{tr}}(x, z), d_{\mathrm{tr}}(x, z)\} = \max\{\max_i(x_i, z_i) - \min_i(x_i - z_i), \max_i(y_i, z_i) - \min_i(y_i - z_i)\}.$$

Also,

$$\begin{aligned} \min_i(x_i - z_i) &\leq \min_i\{\max(x_i - z_i, y_i - z_i)\} \\ \min_i(y_i - z_i) &\leq \min_i\{\max(x_i - z_i, y_i - z_i)\}, \end{aligned}$$

and

$$\max\{\max_i(x_i - z_i), \max_i(y_i - z_i)\} = \max_i\{\max(x_i - z_i, y_i - z_i)\}.$$

Therefore,

$$\begin{aligned} &\max_i\{\max(x_i - z_i, y_i - z_i)\} - \min_i\{\max(x_i - z_i, y_i - z_i)\} \\ &\leq \max\{\max_i(x_i, z_i) - \min_i(x_i - z_i), \max_i(y_i, z_i) - \min_i(y_i - z_i)\}. \end{aligned}$$

Thus,

$$d_{\mathrm{tr}}(x \oplus y, z) \leq \max\{d_{\mathrm{tr}}(x, z), d_{\mathrm{tr}}(x, z)\}.$$

□

Proof for Theorem 3. Let $u, v \in B_x(r)$. Then, we have $d_{\mathrm{tr}}(x, u) \leq r$ and $d_{\mathrm{tr}}(x, v) \leq r$. We want to show any points in the tropical line segment $\Gamma_{u,v}$ between u, v is in $B_x(r)$. Let $z \in \Gamma_{u,v}$. Then there exist $a, b \in \mathbb{R}$ such that

$$z = a \odot u \oplus b \odot v.$$

Thus we have

$$\begin{aligned} d_{\mathrm{tr}}(z, x) &= d_{\mathrm{tr}}(a \odot u \oplus b \odot v, x) \\ &\leq \max\{d_{\mathrm{tr}}(a \odot u, x), d_{\mathrm{tr}}(b \odot v, x)\} \text{by Lemma 3} \\ &= \max\{d_{\mathrm{tr}}(u, x), d_{\mathrm{tr}}(v, x)\} \text{by Lemma 2} \\ &\leq r. \end{aligned}$$

Thus, $z \in B_x(r)$. □

Funding: This research was funded by National Science Foundation DMS 1916037.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The author declare no conflict of interest. The funders had no role in the design of the study.

References

1. Maddison, W.P. Gene trees in species trees. *Syst. Biol.* **1997**, *46*, 523–536. [CrossRef]
2. Huson, D.H.; Klopper, T.; Lockhart, P.J.; Steel, M.A. *Reconstruction of Reticulate Networks from Gene Trees*; Research in Computational Molecular Biology, Proceedings, Springer: Berlin, Germany, 2005; pp. 233–249.
3. Weisrock, D.W.; Shaffer, H.B.; Storz, B.L.; Storz, S.R.; Storz, S.R.; Voss, S.R. Multiple nuclear gene sequences identify phylogenetic species boundaries in the rapidly radiating clade of Mexican ambystomatid salamanders. *Mol. Ecol.* **2006**, *15*, 2489–2503. [CrossRef] [PubMed]
4. Taylor, J.W.; Jacobson, D.J.; Kroken, S.; Kasuga, T.; Geiser, D.M.; Hibbett, D.S.; Fisher, M.C. Phylogenetic species recognition and species concepts in fungi. *Fungal Genet. Biol.* **2000**, *31*, 21–32. [CrossRef] [PubMed]
5. Owen, M.; Yoshida, R. *Continuous Spaces of Phylogenetic Trees*; 2020. in preparation.
6. Speyer, D.; Sturmfels, B. Tropical mathematics. *Math. Mag.* **2009**, *82*, 163–173. [CrossRef]
7. Yoshida, R.; Zhang, L.; Zhang, X. Tropical Principal Component Analysis and its Application to Phylogenetics. *Bull. Math. Biol.* **2019**, *81*, 568–597. [CrossRef] [PubMed]
8. Akian, M.; Gaubert, S.; Viorel, N.; Singer, I. Best approximation in max-plus semimodules. *Linear Algebra Appl.* **2011**, *435*, 3261–3296. [CrossRef]
9. Cohen, G.; Gaubert, S.; Quadrat, J. Duality and separation theorems in idempotent semimodules. *Linear Algebra Appl.* **2004**, *379*, 395–422. [CrossRef]
10. Monod, A.; Lin, B.; Yoshida, R.; Kang, Q. Tropical Geometry of Phylogenetic Tree Space:A Statistical Perspective, 2019. Available online: https://arxiv.org/pdf/1805.12400.pdf (accessed on 9 March 2021).
11. Fix, E.; Hodges, J. *Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties*; Technical Report 4, USAF School of Aviation Medicine, Randolph Field: San Antonio, TX, USA, 1951.
12. Saadatfar, H.; Khosravi, S.; Joloudari, J.H.; Mosavi, A.; Shamshirband, S. A New K-Nearest Neighbors Classifier for Big Data Based on Efficient Data Pruning. *Mathematics* **2020**, *8*, 286. [CrossRef]
13. Terrell, G.R.; Scott, D.W. Variable Kernel Density Estimation. *Ann. Stat.* **1992**, *20*, 1236–1265. [CrossRef]
14. Costa, J.A.; Hero, A.O. Manifold learning using Euclidean k-nearest neighbor graphs [image processing examples]. In Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, QC, Canada, 17–21 May 2004; Volume 3, p. iii–988. [CrossRef]
15. Maclagan, D.; Sturmfels, B. *Introduction to Tropical Geometry; Vol. 161, Graduate Studies in Mathematics*; Graduate Studies in Mathematics, 161, American Mathematical Society: Providence, RI, USA, 2015.
16. Billera, L.; Holmes, S.; Vogtmann, K. Geometry of the space of phylogenetic trees. *Adv. Appl. Math.* **2001**, *27*, 733–767. [CrossRef]
17. Lin, B.; Sturmfels, B.; Tang, X.; Yoshida, R. Convexity in Tree Spaces. *SIAM Discret. Math.* **2017**, *3*, 2015–2038. [CrossRef]
18. Maddison, W.P.; Maddison, D. Mesquite: A Modular System for Evolutionary Analysis. Version 2.72, 2009. Available online: http://mesquiteproject.org (accessed on 9 March 2021).
19. Ripley, B. Package "Class", 2020. Available online: http://www.stats.ox.ac.uk/pub/MASS4/ (accessed on 9 March 2021).