Ordinal and convex assumptions in phylogenetic tree reconstruction

---

A thesis
submitted in partial fulfilment
of the requirements for the Degree
of
Master of Science in Mathematics
in the
University of Canterbury
by
Robin P. Candy

University of Canterbury
2014

Contents

Acknowledgements

First and foremost, I wish to express my deep gratitude to my wonderful supervisors Charles Semple and Mike Steel. They both put a lot of effort into keeping me on track and making sure I was attempting work possible in the given time constraints. I am also thankful for the support received from the School of Mathematics and Statistics, in terms of study space and resources in a turbulent time (the whole department was covered in scaffolding and many areas were building sites).

For the financial assistance I received during the course of this thesis, I am thankful to both the University of Canterbury and the New Zealand Marsden Fund.

I thank Amanda Deacon for her helpful comments and advice in early drafts. Finally, I thank Boukje Breedvelt, Barry Candy, Tim Candy and Yvonne Candy for their support.

Abstract

Phylogenetics is a field primarily concerned with the reconstruction of the evolutionary history of present day species. Evolutionary history is often modeled by a phylogenetic tree, similar to a family tree. To recreate a phylogenetic tree from information about current species, one needs to make assumptions about the evolutionary process. These assumptions can range from full parametrised models of evolution to simple observations. This thesis looks at the reconstruction of phylogenetic trees under two different assumptions. The first, known as the ordinal assumption, has been previously studied and asserts that as species evolve, they become more dissimilar. The second, the convex assumption, has not previously been studied in this context and asserts that changes species go through to become dissimilar are progressively larger than the current differences between those species.

This thesis presents an overview of mathematical results in tree reconstruction from dissimilarity maps (also known as distance matrices) and develops techniques for reasoning about the ordinal and convex assumptions. In particular, three main results are presented: a complete classification of phylogenetic trees with four leaves under the ordinal assumption; a partial classification of phylogenetic trees with four leaves under the convex assumption; and, an independent proof of a result on the relationship between ultrametrics and the ordinal assumption.

Chapter 1

Introduction

Phylogenetics is the study of evolutionary relationships. These relationships are often represented by phylogenetic X-trees, in which leaves represent extant entities from a finite set X (e.g. species existing today), internal vertices represent extinct hypothetical entities and edges represent evolutionary connection (e.g. ancestry). A core problem in phylogenetics is the following: given information about extant entities, how does one construct a phylogenetic X-tree, and does the tree constructed represent the true evolutionary history for those entities? In order to address the second part of this problem, assumptions about the true evolutionary process must be made. If correct, these assumptions (along with the information on the extant entities) narrow the set of possible phylogenetic X-trees to just those that can represent the true evolutionary history of the initial entities. Ideally, the assumptions made are strict enough so that exactly one phylogenetic X-tree is identified. This thesis is concerned with studying the phylogenetic X-tree construction problem under the ordinal and convex assumptions (defined in Sections 1.3.1 and 1.3.2), with emphasis placed on cases with four extant entities.

This thesis is organised as follows. The rest of this chapter is devoted to giving a brief non-technical introduction to the relevant phylogenetic concepts and the inspiration for this thesis. Of particular importance is Section 1.3 on evolutionary assumptions, as this is essentially the sole motivation for this thesis and justifies the importance of the two assumptions studied.

Chapter 2 gives an introduction to graph theory and formally introduces X-trees (Section 2.3), of which phylogenetic X-trees are a specialisation. A special type of phylogenetic X-tree known as a quartet is studied in Section 2.4. In Chapter 3, a formal treatment of the assumptions studied is given and, in doing so, the connection between dissimilarity maps (Section 3.3), metrics (Section 3.4) and phylogenetic X-trees is analysed (Section 3.5). Chapter 4 extends the order interval relation, a tool used in Kearney (1998) for reasoning about the ordinal assumption, to a new convex interval relation (Section 4.2) for reasoning about the convex assumption. Various properties and examples of the two interval relations are given. The final chapter (Chapter 5) classifies an arbitrary dissimilarity map δ on four entities, under both the ordinal (Section 5.2) and convex (Section 5.3) assumptions, by explicitly finding all phylogenetic X-trees associated with δ. The classification is aided by the interval relations from Chapter 4, which take the sting out of many of the classification proofs. Chapter 5 concludes with an independent proof of a known result based on Proposition 7.5.6. in Semple and Steel (2003) pertaining to the relationship between the ordinal assumption and ultrametrics.

All proofs and numbered examples contained in this thesis are formulated independently of other sources. Non-trivial results which, to the author's knowledge, have not been previously published appear in Sections 5.2 and 5.3. To the author's knowledge, the convex assumption has not previously been studied in the context of phylogenetic X-tree reconstruction. The main resource for the notation and definitions in this thesis is Semple and Steel (2003).

1.1 History

Phylogenetics is mainly motivated by biology, in which constructing the evolutionary history of sets of species is of keen interest. Figure 1.1, from Darwin (1837), shows one of the earliest examples of a phylogenetic X-tree used to describe evolutionary history. That phylogenetic X-tree (and subsequent other early phylogenetic X-trees) are based on similarities between the physical characteristics of different species. Such characteristics can be

misleading, as similar characteristics do not necessarily imply an evolutionary connection (e.g. a close common ancestor) and vice versa. For example, environmental pressures can force similar traits to emerge independently in unrelated species, as in the case of bats and birds both being able to fly.

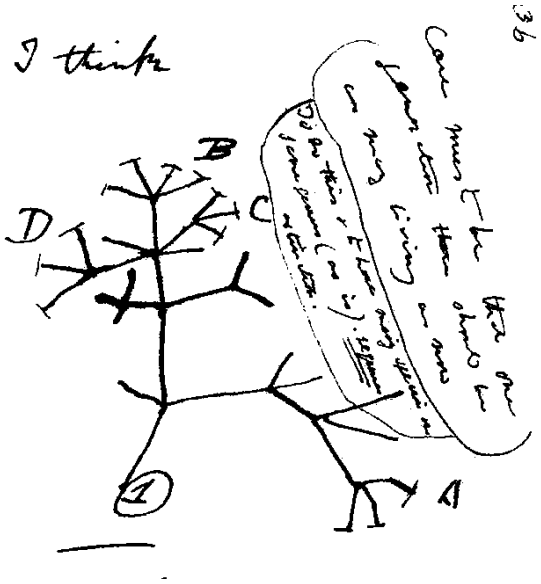


Figure 1.1: Charles Darwin's famous example of a graph used to convey evolutionary relationships between species.

With the advent of protein and genetic sequencers by Edman and Begg (1967) and Sanger et al. (1977) (respectively), comparisons between species shifted from being based on characteristics to being based on molecular information. This allowed for more accurate phylogenetic X-tree reconstruction and led to many new discoveries in the understanding of our evolutionary history. Molecular comparisons also necessitated new discrete mathematics, statistical techniques and efficient reconstruction algorithms be developed to accurately deal with the large amount of new data. For this reason, modern phylogenetics represents a melting pot of ideas from biology, discrete mathematics, statistics and computer science.

1.2 Evolutionary distance

There are two main types of data used to reconstruct phylogenetic X-trees: characters and dissimilarity maps. Characters are comparable descriptions of each entity's attributes (e.g. DNA sequence data or whether a species can fly). Dissimilarity maps are numerical descriptions of the differences between each pair of entities (e.g. average beak length difference in millimetres between species of bird). Characters can be used to construct dissimilarity maps (but not generally vice versa); in this sense characters are a more general data type. In this thesis, we are interested in reconstruction based on dissimilarity maps, as this assumes that the context-dependent task of comparing characters in a meaningful way has already been done.

True evolutionary distance between entities may represent a number of different factors. Some examples are as follows: time since common ancestry; number of evolutionary events since common ancestry (e.g. combined number of speciation events in both species' evolutionary histories); or, in a biological example, number of DNA sequence changes since speciation from a common ancestor. Finding the true evolutionary distance between each entity is analogous to finding the (weighted) phylogenetic X-tree which correctly describes the evolutionary history of the set of entities X. Dissimilarity maps can be thought of as an encoding of observed evolutionary distance. Observed evolutionary distance is any quantity that may be indicative of true evolutionary distance (i.e. any quantity that is affected by the same evolutionary process). For example, Buneman (a pioneer of mathematical phylogenetics) looked at the recovery of document copy history from transcription errors. In his paper, Buneman (1971), he described true evolutionary distance as the number of letter changes between two documents since being copied. Similarly, he describes the observed evolutionary distance as the number of discrepancies between copies.

1.3 Evolutionary assumptions

As discussed in the previous section, one type of data used to reconstruct phylogenetic X-trees is based on dissimilarity maps, in which observed evo-

lutionary distances between pairs of entities are tabulated. The common assumption in all reconstruction methods[1] is that observed evolutionary distance is indicative of true evolutionary distance. The exact nature of how the two quantities are related depends on the context of the evolutionary process, and gives rise to many different assumptions and corresponding reconstruction techniques.

If a dissimilarity map encodes true evolutionary distance and not just observed evolutionary distance, it is called *additive* (see Kearney (1998) for a more formal definition). One of the most popular assumptions in reconstruction methods is the assumption of additivity, in which a dissimilarity map is assumed to be close to additive (i.e. true evolutionary distance is assumed to be approximated by observed evolutionary distance). Popular reconstruction methods that employ this assumption (or a stricter version of it) include neighbour-joining (Saitou and Nei, 1987), split decomposition (Bandelt and Dress, 1992) and UPGMA (Sokal and Sneath, 1963). A key difficulty in reconstruction under this assumption is that dissimilarity maps do not necessarily reflect the evolutionary changes the entities went through to become dissimilar accurately enough. For example, suppose we wish to recover the evolutionary history (i.e. a phylogenetic X-tree) for a set of species X using DNA sequence comparisons. At some time in history, all the species in X had the same DNA, but the individual sites along each species DNA changed as the species evolved. The true evolutionary distance in this example is the total number of site changes that have occurred (added together) in two species' DNA as it mutated from their closest common ancestor's DNA. The observed distance is the number of site discrepancies between two species' DNA. In this example, the true evolutionary distance is being progressively underestimated by observed evolutionary distance, as site discrepancies do not indicate how many site changes have occurred since speciation.

To make the assumption of additivity more appropriate, many methods

[1]When reconstruction techniques and methods are brought up in this thesis, we are referring to reconstruction of phylogenetic X-trees from dissimilarity maps on a set of entities X. This is different to reconstruction from character data, which has a distinct set of reconstruction methods such as maximum parsimony (Fitch, 1971).

correct the dissimilarity data before reconstruction. Typically, this involves modelling the evolutionary process to counteract the progressive underestimation of true evolutionary distance (i.e. using a transform based on an evolutionary model to correct the dissimilarity map). Under ideal conditions (i.e. the model of evolution closely reflects the evolutionary process), reconstruction using the corrected dissimilarity maps performs better than without (Kearney, 1998); however, correcting dissimilarity data before the reconstruction process has some shortfalls. For instance, an intrinsic understanding of the actual evolutionary process is required to commit to a particular model and its parameters. This presents a problem as understanding of the actual evolutionary process is often lacking. In fact, not knowing the actual evolutionary process well is usually the very motivation for the phylogenetic X-tree reconstruction in the first place.

The next two assumptions presented are explored thoroughly in this thesis. They provide alternatives to the assumption of additivity on corrected dissimilarity maps. The second presented assumption has not previously been studied in the context of tree reconstruction.

1.3.1 The ordinal assumption

A number of authors have looked at the ordinal assumption in tree reconstruction (Guénoche, 1998; Kearney, 1997; Bonnot et al., 1996; Kannan and Warnow, 1993). Kearney classifies it as an assumption about pairs of entities being more or less similar than another pair of entities. For the purpose of this thesis, the ordinal assumption is that the order of a non-decreasing sequence of all true evolutionary distances between entities is the same as the order of a non-decreasing sequence of all observed evolutionary distances between entities. More formally, the ordinal assumption is that there is a strictly-increasing function bringing observed evolutionary distance to real evolutionary distance (Definition 3.23). In terms of DNA, the ordinal assumption is the following implication: discrepancies between DNA sequence A and B are greater than that between A and C, hence the true evolutionary distance between A and B is greater than that between A and C.

Transforms based on the Jukes and Cantor (1969) model of evolution,

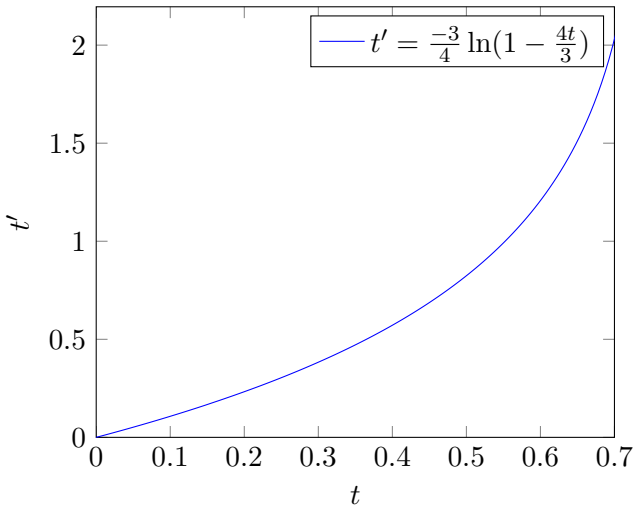


Figure 1.2: Plot showing a sample transform consistent with the Jukes-Cantor model of evolution. Transform is given by $t' = \frac{-3}{4}\ln(1 - \frac{4t}{3})$ where t is the original distance and t' is the corrected distance.

shown in Figure 1.2, support the ordinal assumption. That is, transforms based on the Jukes-Cantor model of evolution are strictly-increasing. Indeed, other models of (DNA) evolution implicitly support the ordinal assumption also (Felsenstein, 1981; Tamura, 1992). The advantage of the ordinal assumption over the assumption of additivity, on corrected dissimilarity maps, is that there is no need to commit to a particular model (or model parameters) for reconstruction. Furthermore, the results based on the ordinal assumption are generally less sensitive to small changes in observed evolutionary distance, as small perturbations leave ordering intact.

In this thesis we present two main results based on the ordinal assumption. The first is a complete classification of dissimilarity maps with four entities (Section 5.2). This work is based on Kearney et al. (1999), which uses a similar partial classification to justify a quartet classification algorithm based on the ordinal assumption. The results in Section 5.2 are not previously published. The second result (presented in Section 5.5) is an independent proof a known result related to Proposition 7.5.6. from Semple and Steel (2003). The result relates to the relationship between ultrametrics and the ordinal assumption.

1.3.2 The convex assumption

The convex assumption is closely related to the ordinal assumption. The convex assumption is that there is a strictly-increasing convex function bringing observed evolutionary distance to real evolutionary distance (Definition 3.24). The convex assumption is appropriate when there is progressive underestimation of real evolutionary distance from dissimilarity data. From a molecular biology point of view, the convex assumption is justified by 'hidden' mutation, in which sites on sequences of DNA have evolved multiple times or evolved back to their original states. It is in this sense that observed evolutionary distance can progressively underestimate true evolutionary distance, as true evolutionary distance takes into account hidden mutations and observed evolutionary distance does not. In this case, the convexity assumption is asserting that the higher the number of discrepancies between sites on sequences A and B, the more hidden mutation events have occurred on both A and B since speciation from a common ancestor sequence.

A similar justification to that of the ordinal assumption can be provided for the convex assumption. That is, all the models of evolution presented in the previous section that support the ordinal assumption also support the convex assumption. That is, correction transforms based on these models are not only strictly-increasing but also convex. To understand why, consider the fact that DNA sequence evolutionary models are based on individual sites randomly changing under some process (usually a Markov process). As long as the process is time-independent (i.e. sites change regardless of previous site changes), there is a correlation between number of site changes and number of hidden mutation events. As was established in the previous paragraph, that is essentially all that is needed to support the convex assumption.

The reconstruction problem under the convex assumption has not previously been studied. In this thesis, we provide a formal framework for reasoning about the convex assumption and, within this framework, present a partial classification of dissimilarity maps consisting of four entities (Section 5.3).

Chapter 2

Graphs and X-trees

Graphs are a misleadingly simple structure. They are simple in the sense that they have very few restrictions and can be very concisely conveyed. However, graphs are also very complex as, despite many deep results in graph theory (Robertson and Seymour, 1985), basic questions about their structure have yet to be answered. Fundamental questions such as the Kelly-Ulam Reconstruction Conjecture (see Bondy and Hemminger (1977) for details), which asks whether it is possible to find a graph given all its subgraphs on one less vertex remain unsolved.

Besides being studied for their intrinsic beauty, many of the problems in graph theory have physical inspiration. Graphs have long been used as a fundamental way to represent relationships in the real world. In physics, they are used to reason about forces applied to objects. In chemistry, they can represent atoms and bonds. From decisions in turn-based games to the commonalities between ancient languages, graphs have proved a useful tool in visualisation, reasoning and representation. Biology has similarly embraced graphs. Graphs, or more specifically phylogenetic X-trees, have been used to represent evolutionary relationships between species since Charles Darwin first proposed the Theory of Evolution (see Figure 1.1).

This chapter explores some basic definitions and results in graph theory in relation to phylogenetic X-trees. In Sections 2.1 and 2.2, graphs and trees are formally introduced. Phylogenetic X-trees are presented in Section 2.3, with special emphasis placed on the importance of quartets (Section 2.4).

In short, this chapter introduces the structures studied in this thesis.

2.1 Graphs

Definition 2.1. A *graph* $G = (V, E)$ is an ordered pair where the first element is a non-empty set V and the second element is a multiset $E \subseteq \{\{u, v\} : u, v \in V\}$. Elements of V are called *vertices*. Elements of E are *edges*. If not explicitly defined, $V(G)$ and $E(G)$ denote the set of all vertices and edges respectively. If $u, v \in V$ and $e = \{u, v\} \in E$, the vertices u and v are *adjacent* and e is *incident* with u and v. A *loop* is an edge of the form $e = \{u, u\}$ for some $u \in V$.

In this thesis, graphs are often represented by diagrams to avoid large and unreadable sets. Dots in graph diagrams represent vertices and lines between dots represent edges.

Example 2.2. Let G be a graph with $V(G) = \{v_0, v_1, v_2, v_3, v_4, v_5\}$ and $E(G) = \{e_0 = \{v_0, v_1\}, e_1 = \{v_0, v_2\}, e_2 = \{v_0, v_3\}, e_3 = \{v_0, v_4\}, e_4 = \{v_1, v_2\}, e_5 = \{v_3, v_4\}\}$ depicted in Figure 2.1. The vertices v_0 and v_1 are adjacent. As there is no edge containing v_5 and v_4, they are not adjacent. The edge e_4 is incident with v_1 but not incident with v_0.

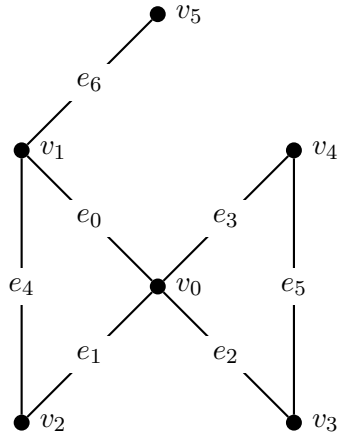


Figure 2.1: A diagram representing the graph G.

Example 2.3. The graph $K_{3,3}$ is depicted in Figure 2.2 and is a *complete bipartite* graph. The vertices of a complete bipartite graph can be split into

two sets A and B such that each vertex in A is adjacent to each vertex in B but no vertex in A or B is adjacent to a vertex in the same set. This particular complete bipartite graph is denoted $K_{3,3}$ as sets $A = \{v_0, v_2, v_4\}$ and $B = \{v_1, v_3, v_5\}$ are both of size three.

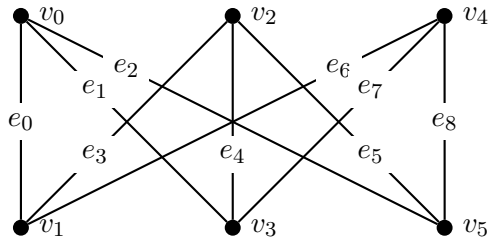


Figure 2.2: The complete bipartite graph $K_{3,3}$.

Definition 2.4. The *degree* of a vertex $v \in V(G)$ in a graph G is the number of non-loop edges incident with it. An edge $e \in E(G)$ is a *pendant edge* if it is incident to a vertex of degree one.

Definition 2.5. A *walk* from v_0 to v_n in a graph $G = (V, E)$ is an alternating list

$$(v_0, e_0, v_1, e_1, v_2 \dots e_{n-1}, v_n)$$

of vertices and edges such that $n \in \mathbb{Z}^+$, $v_0, v_1, v_2 \dots v_n \in V$, $e_0, e_1 \dots e_{n-1} \in E$ and for all i such that $0 \leq i \leq n-1$, $e_i = \{v_i, v_{i+1}\}$. A *path* in G is a walk in which no vertex is contained more than once. A *cycle* in G is a walk in which there are no repeated edges and the first and last vertices are the only vertices that are the same.

Example 2.6. Consider the complete bipartite graph $K_{3,3}$ depicted in Figure 2.2. Each vertex has degree three so there are no pendant edges.

The list $(v_0, e_1, v_3, e_1, v_0)$ is a walk as e_1 is incident to v_0 and v_3. However, it is not a cycle as the edge e_1 is contained in the walk multiple times. It is also not a path as v_0 appears twice in the list. Appending the edge e_4 and the vertex v_3 to the list would no longer make it a walk as e_4 is not incident to v_0.

The walk $(v_0, e_1, v_3, e_7, v_4, e_6, v_1)$ is a path as no vertex is contained multiple times. Appending the edge e_0 and the vertex v_0 to the walk forms a cycle.

2.2 Trees

Definition 2.7. A graph is *acyclic* if it contains no cycles. A graph is *connected* if every pair of vertices have a path between them. A *tree* T is a connected and acyclic graph. A *leaf* (plural: *leaves*) in a tree $T = (V, E)$ is a vertex $v \in V$ with degree one or degree zero (in the case where T consists of a single vertex). Every vertex that is not a leaf is an *internal vertex*. A tree $T = (V, E)$ is *binary* if each vertex $v \in V$ is either a leaf or of degree three.

For brevity the vertex or edge sets of a graph are not explicitly defined unless individual vertices or edges are referred to. When the individual elements of the edge or vertex sets are not made explicit, the labelling of them on a graph diagram is similarly omitted.

Example 2.8. Consider the graph T shown in Figure 2.3. The graph is connected and has no cycles. Hence, the graph T is a tree. As each internal vertex (v_1, v_2 and v_5) is of degree three, the tree T is binary. The tree T has five leaves.

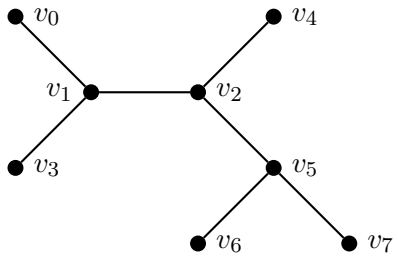
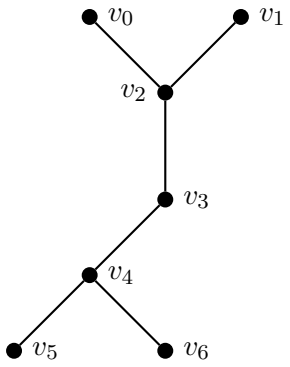


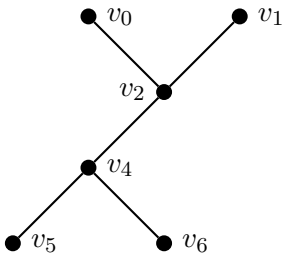
Figure 2.3: A binary tree with eight vertices.

Internal vertices of degree two in trees are sometimes ignored. If the vertices of a graph are not explicitly relevant, then the internal vertices of degree two add no structure to a tree. If this is the case, one may wish just to omit or remove them. More formally, let T be a tree and v be an internal vertex of degree two. A new tree T' can be obtained by removing v from the vertex set and replacing its two incident edges with one edge connecting the adjacent vertices of v. This is known a *suppressing* a vertex. In this case, T' was obtained from T by suppressing v.

Example 2.9. Consider the tree T_1 depicted in Figure 2.4. It has four leaves and three internal vertices. The vertex v_3 is an internal vertex of degree two. Suppressing v_3 gives the tree T_2 (also depicted in Figure 2.4).



(a) The tree T_1.



(b) The tree T_2.

Figure 2.4: The tree T_2 is obtained from T_1 by suppressing the vertex v_3.

Sometimes, real valued positive weights are assigned to each edge in a tree using a *weight function*. A tree and a consistent weight function pair are referred to as a *weighted tree*. This is useful as it allows us to display comparative distance between vertices. For instance, if vertices on a tree represent species and edges represent ancestry, then a weight function could be used to indicate the time between speciation events.

Example 2.10. Consider the tree S and weight function $w : E(S) \to \mathbb{R}^+$ shown in Figure 2.5. It is a *star tree*. Star trees are defined by the property of having only one internal vertex. The degree of this vertex is eight, so it is not binary. The pair (S, w) is a weighted tree.

Often in graph theory, when depicting a weighted tree, different edge weights are given proportional edge lengths. This convention is not followed in this thesis as it can lead to confusion on how a weight function and a graph are related; a weight function does not modify or warp a graph's edges. Also, to prevent ambiguity between edge labels and weights in graph diagrams, edges are never labelled on weighted trees.

Definition 2.11. Two trees $T_1 = (V_1, E_1)$ and $T_2 = (V_2, E_2)$ are *isomorphic* if there exists a bijection $\psi : V_1 \to V_2$ such that, for all veritices $u, v \in V_1$,

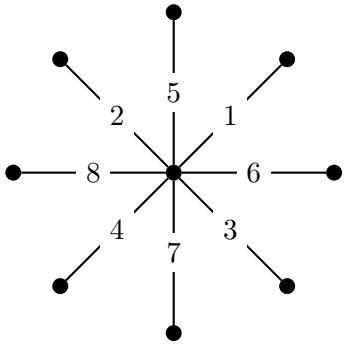
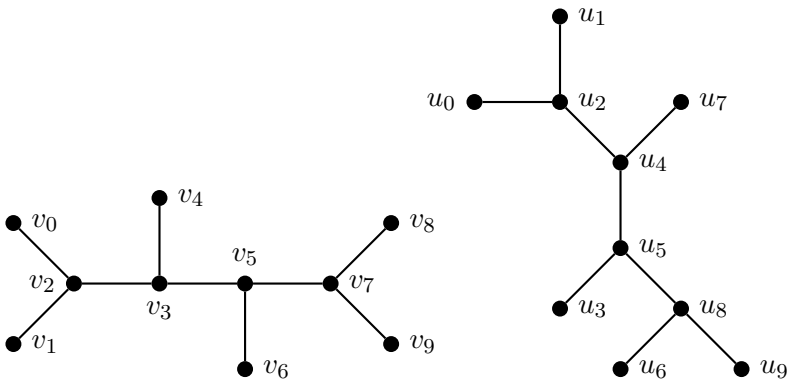


Figure 2.5: A weighted star tree with eight pendant edges.

$\{u, v\} \in E_1$ if and only if $\{\psi(u), \psi(v)\} \in E_2$. In this case ψ is known as an *isomorphism* between T_1 and T_2.

Example 2.12. The binary trees T_1 and T_2 (depicted in Figure 2.6) are isomorphic. There are multiple isomorphisms from T_1 to T_2. One such isomorphism is $\psi : V(T_1) \to V(T_2)$ with

$$\psi = \begin{array}{c|cccccccccc} V(T_1) & v_0 & v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & v_7 & v_8 & v_9 \\ \hline \psi(V(T_1)) & u_0 & u_1 & u_2 & u_4 & u_7 & u_5 & u_3 & u_8 & u_9 & u_6 \end{array}.$$



(a) The binary tree T_1. (b) The binary tree T_2.

Figure 2.6: Two isomorphic binary trees.

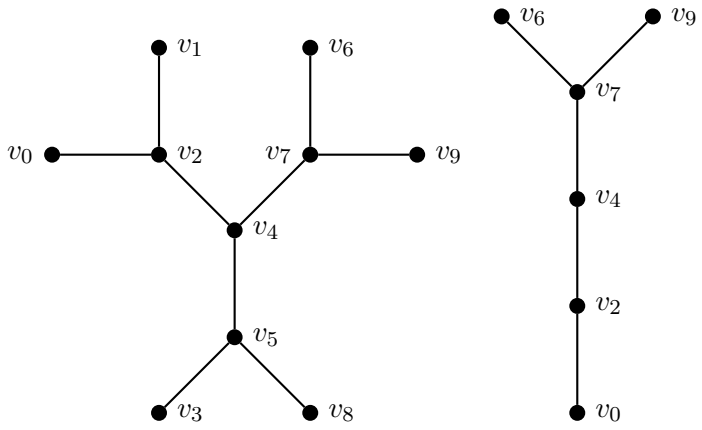
To be able to reason about the similarities between large trees and smaller trees on the same vertex set, the following definition is introduced.

It gives a way of restricting a tree on some set of vertices to a smaller tree on a subset of those vertices.

Definition 2.13. Let $T = (V, E)$ be a tree, $U \subset V$ and $P_T(v_0, v_1)$ (with $v_0, v_1 \in V$) denote the (unique) path between vertices v_0 and v_1 in T. Consider the tree T' with $V(T') = \{v \in P_T(u_0, u_1) : u_0, u_1 \in U\}$ and $E(T') = \{e \in P_T(u_0, u_1) : u_0, u_1 \in U\}$. The tree T' is the tree T *restricted* to U and is denoted $T|U$. The tree $T|U$ is also referred to as the (U-)*restricted subtree* of T.

Note that different subsets of the vertex set can result in the same restricted subtrees of a tree.

Example 2.14. Consider the binary tree T shown in Figure 2.7a and the set $V' = \{v_0, v_4, v_7, v_6, v_9\}$. The restricted subtree $T|V'$ contains all the edges and vertices on paths between vertices in V'. As the path from v_0 to v_6 contains v_2, v_4 and v_7, the restricted subtree $T|V' = T|\{v_0, v_6, v_9\}$. No path between vertices in V' contains v_1, v_3, v_5 or v_8. Therefore $V(T|V') = \{v_0, v_2, v_4, v_6, v_7, v_9\}$ and $T|V'$ is the tree displayed in Figure 2.7b.



(a) The binary tree T. (b) The tree $T|\{v_0, v_6, v_9\}$.

Figure 2.7: The tree T and its restricted subtree $T|v_0, v_6, v_9$.

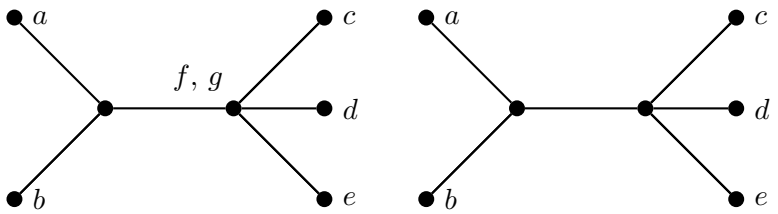
2.3 Phylogenetic X-trees

Definition 2.15. An *X-tree* $\mathcal{T}$ is an ordered pair (T, ϕ), where $T = (V, E)$ is a tree and $\phi : X \to V$ is a function such that, for each leaf $l \in V$ and vertex of degree two $v_2 \in V$, both $l, v_2 \in \phi(X)$. If, in addition, ϕ is a bijection on the leaves of T, then $\mathcal{T}$ is a *phylogenetic X-tree*. The tree T is the *underlying tree* of $\mathcal{T}$ and ϕ is the *labelling function* of $\mathcal{T}$. If T is not explicitly defined, $T(\mathcal{T})$ is the underlying tree. Similarly, $\phi(\mathcal{T})$ is the labelling function.

As with regular trees, an X-tree and weight function pair is referred to as a weighted X-tree. A diagram for an X-tree $\mathcal{T} = (T, \phi)$ displays elements of X next to a vertex as dictated by ϕ. In this sense ϕ can be thought of as a labelling function. To prevent conflict between vertex labels and elements of X, the diagram of the X-tree $\mathcal{T}$ never includes labels of the vertices of the underlying tree T. If the underlying tree of an X-tree $\mathcal{T}$ is binary, then $\mathcal{T}$ is a binary X-tree. The notation $V(\mathcal{T})$ and $E(\mathcal{T})$ is used to refer to the vertices and edges of the underlying tree, respectively.

Example 2.16. Let $X = \{a, b, c, d, e, f, g\}$. Consider the X-tree $\mathcal{T}$ shown in Figure 2.8a. It is not binary and is not a phylogenetic X-tree as $\phi(\mathcal{T})$ is not a bijection on the leaves of the underlying tree $T(\mathcal{T})$. It is an X-tree as each leaf of the underlying tree $T(\mathcal{T})$ is in the range of $\phi(\mathcal{T})$.

Example 2.17. Let $X = \{a, b, c, d, e\}$. Consider the X-tree $\mathcal{T}$ shown in Figure 2.8b. It is not binary but is a phylogenetic X-tree as $\phi(\mathcal{T})$ is a bijection on the leaves of the underlying tree $T(\mathcal{T})$.



(a) A X-tree on the set $X = \{a, b, c, d, e, f, g\}$.

(b) A phylogenetic X-tree on $X = \{a, b, c, d, e\}$.

Figure 2.8: Examples of both a X-tree and a phylogenetic X-tree.

Definition 2.18. Two X-trees $\mathcal{T}_1 = (T_1, \phi_1)$ and $\mathcal{T}_2 = (T_2, \phi_2)$ are *isomorphic* if there exists an isomorphism, $\psi : V_1 \to V_2$ between $T_1 = (V_1, E_1)$ and $T_2 = (V_2, E_2)$ such that $\phi_2 = \psi \circ \phi_1$.

Example 2.19. Let $X = \{a, b, c, d, e, f\}$. Consider the binary phylogenetic X-trees $\mathcal{T}_1 = (T_1, \phi_1)$ and $\mathcal{T}_2 = (T_2, \phi_2)$ (depicted in Figure 2.9) with base trees from Example 2.12. The labelling functions for $\mathcal{T}_1$ and $\mathcal{T}_2$ are

$$\phi_1 = \begin{array}{c|cccccc} X & a & b & c & d & e & f \\ \hline \phi_1(X) & u_0 & u_1 & u_7 & u_3 & u_9 & u_6 \end{array}$$

and

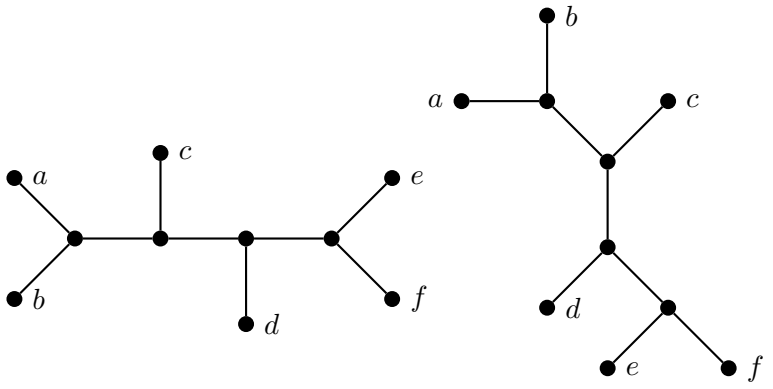
$$\phi_2 = \begin{array}{c|cccccc} X & a & b & c & d & e & f \\ \hline \phi_1(X) & v_0 & v_1 & v_4 & v_6 & v_8 & v_9 \end{array},$$

respectively.

The tree isomorphism $\psi : V(T_1) \to V(T_2)$ with

$$\psi = \begin{array}{c|cccccccccc} V(T_1) & v_0 & v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & v_7 & v_8 & v_9 \\ \hline \psi(V(T_1)) & u_0 & u_1 & u_2 & u_4 & u_7 & u_5 & u_3 & u_8 & u_9 & u_6 \end{array},$$

from Example 2.12, is also an isomorphism between the binary phylogenetic X-trees $\mathcal{T}_1$ and $\mathcal{T}_2$ as $\phi_2 = \psi \circ \phi_1$. Hence, $\mathcal{T}_1$ and $\mathcal{T}_2$ are isomorphic.



(a) The binary phylogenetic X-tree $\mathcal{T}_1$.

(b) The binary phylogenetic X-tree $\mathcal{T}_2$.

Figure 2.9: Two isomorphic binary phylogenetic X-trees.

In this thesis, a X-tree is considered equivalent to all X-trees isomorphic with it. For formal consistency, consider an X-tree as the isomorphism class of that X-tree.

Proposition 2.20 gives an idea of the combinatorial explosion in the number of phylogenetic X-trees as X increases in size ($b(3) = 1$, $b(4) = 3$, $b(5) = 15$, $b(10) = 2027025$, $b(15) = 7905853580625\ldots$). It is desirable to know which binary phylogenetic X-tree of some size is the closest match to some set data. The following proposition highlights the infeasibility of checking all phylogenetic X-trees of a given size. Similarly, it also gives us an indication for the sizes of X where a brute-force search is feasible. A proof of the result is in Semple and Steel (2003, p. 17).

Proposition 2.20. *Let $n \in \mathbb{Z}^+$ such that $n > 2$, the set X have $|X| = n$ and $b(n)$ be the number of distinct binary phylogenetic X-trees. Then,*

$$b(n) = \frac{(2n-4)!}{(n-2)!2^{n-2}}.$$

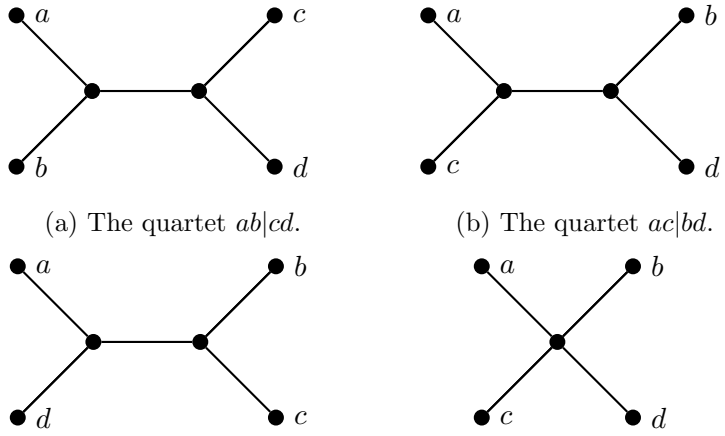
2.4 Quartets

Phylogenetic X-trees in which the size of X is four play an important role in the reconstruction of larger trees. A phylogenetic X-tree is uniquely defined by the structure of some subsets of X. Many paradigms for reconstructing phylogenetic X-trees rely on this to create a two-step reconstruction process: First, they generate a set of smaller phylogenetic X'-trees from some data about the desired phylogenetic X-tree. Second, they combine the smaller phylogenetic X'-trees to reconstruct the original phylogenetic X-tree. This paradigm is followed in the Quartet Puzzling method of Strimmer and Von Haeseler (1996) and, more recently the Ordinal Quartet (Kearney, 1998), Rec-I-DCM3 (Roshan et al., 2004) and Short Quartet Puzzling (Snir et al., 2008) reconstruction methods. This section focuses on phylogenetic X'-trees (with $|X'| = 4$) and how they can be used in the reconstruction of larger phylogenetic X-trees.

Definition 2.21. A *phylogenetic 4-tree* $\mathcal{T}$ on X is a phylogenetic X-tree with $|X| = 4$. If $\mathcal{T}$ is also binary, then $\mathcal{T}$ is a *quartet*.

A quartet on $X = \{a, b, c, d\}$ is often notated $ab|cd$, $ac|bd$ or $ad|bc$. The partitions of X either side of the | symbol refer to sets of leaves adjacent to the same (internal) vertex. Another, more general, way of thinking of this is that the paths between vertices on the left of the partition do not intersect path between vertices on the right of the partition. In a similar way, the star tree on X can be notated $abcd$ as each leaf is adjacent to the same internal vertex.

Example 2.22. The phylogenetic 4-trees $ab|cd$, $ac|bd$, $ad|bc$ and $abcd$ on $\{a, b, c, d\}$ are displayed in Figure 2.10.



(a) The quartet $ab|cd$.

(b) The quartet $ac|bd$.

(c) The quartet $ad|bc$.

(d) The star tree $abcd$

Figure 2.10: All phylogenetic 4-trees on the set $\{a, b, c, d\}$.

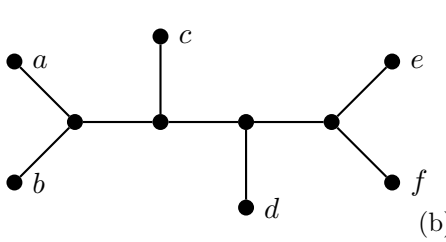
The phylogenetic 4-trees from Example 2.22 are the only possible phylogenetic 4-trees. This is shown in Proposition 2.23.

Proposition 2.23. *Let $X = \{a, b, c, d\}$. A phylogenetic X-tree $\mathcal{T}$ is exactly one of the trees $ab|cd$, $ac|bd$, $ad|bc$ or $abcd$.*

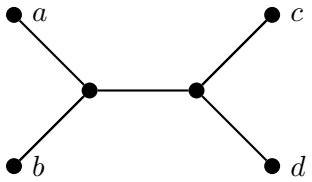
Proof. By Proposition 2.20 there are just three quartets. Therefore if $\mathcal{T}$ is binary it is one of $ab|cd$, $ac|bd$ or $ad|bc$ (these are non isomorphic as they partition leaf labels differently either side of the edge incident with both internal vertices). Assume $\mathcal{T}$ is not binary. Then, $T(\mathcal{T})$ contains an internal vertex with degree greater than three. But as a phylogenetic X-tree has no internal vertices of size two and each leaf has a unique label, $\mathcal{T} = abcd$. □

Definition 2.24. Let $\mathcal{T} = (T, \phi)$ be an X-tree and $X' \subset X$. Let T' be the restricted tree $T|\phi(X')$ with all vertices of degree two not in $\phi(X')$ suppressed. Let $\phi|X'$ denote the restriction of the function ϕ to the domain X' and $\phi' = \phi|X'$. The X'-tree $\mathcal{T}' = (T', \phi')$ is the (X'-)*restricted phylogenetic subtree* of $\mathcal{T}$.

Example 2.25. Consider the X-tree $\mathcal{T}$ depicted in Figure 2.11a with $X = \{a, b, c, d, e, f\}$. Restricting $\mathcal{T}$ to $X' = \{a, b, c, d\}$ gives the X'-tree $\mathcal{T}|X'$ shown in Figure 2.11b.



(a) The phylogenetic X-tree $\mathcal{T}$.



(b) The phylogenetic X'-tree $\mathcal{T}|\{a, b, c, d\}$.

Figure 2.11: Depiction of the phylogenetic X-tree $\mathcal{T}$ and $\mathcal{T}|\{a, b, c, d\}$. Note the suppression of the internal vertices of degree two (on the path from c to d).

The following theorem shows why phylogenetic 4-trees are so useful. Given the set of all restricted phylogenetic 4-trees of some phylogenetic X-tree $\mathcal{T}$, one can uniquely reconstruct $\mathcal{T}$. A proof of this result is provided in Semple and Steel (2003, p. 117).

Theorem 2.26. *Let $\mathcal{T}$ and $\mathcal{T}'$ be phylogenetic X-trees with $|X| \geq 4$. Let $\mathcal{Q}(\mathcal{T})$ and $\mathcal{Q}(\mathcal{T}')$ be the sets of all phylogenetic 4-trees that are restricted subtrees of $\mathcal{T}$ and $\mathcal{T}'$, respectively. If $\mathcal{Q}(\mathcal{T}) = \mathcal{Q}(\mathcal{T}')$, then $\mathcal{T} = \mathcal{T}'$.*

Chapter 3

Functions and tree metrics

The distances between leaves on a weighted phylogenetic X-tree are important as they encode evolutionary distance. When tabulated, these distances form a strict tree metric (Definition 3.14). A phylogenetic X-tree is uniquely defined by its corresponding strict tree metric (Theorem 3.16). The process of transforming dissimilarity maps (representing observed evolutionary distance) to find a consistent tree metric (representing true evolutionary distance) is very useful in a biological setting. If the function used to transform a dissimilarity map is consistent with the evolutionary process, the resulting phylogenetic X-tree represents the reconstruction of a possible evolutionary history for the set of species X. The higher the confidence that the function matches the true evolutionary process, the more likely the phylogenetic X-tree found will represent true evolutionary history. Convex and order-preserving functions are studied as they are the basis of the convex and ordinal assumptions (respectively). Ideally, the set of candidate tree metrics for a given dissimilarity map is non-empty and very small.

This chapter introduces order-preserving functions and some useful and defining properties of convex functions (Sections 3.1 and 3.2). These basic definitions and properties are relied on heavily, as they characterise the transformations supported by the ordinal and convex assumptions. The concepts and properties of dissimilarity maps and tree metrics are formally introduced in Sections 3.3 and 3.4. This chapter concludes with Section 3.5, which ties together the concepts of convex and order-preserving functions,

X-trees and dissimilarity maps.

The main resource for the notation and definitions used in this chapter is Semple and Steel (2003).

3.1 Order-preserving functions

In the first type of function introduced (Definition 3.1), the order of elements in the domain of a function is the same order as of the corresponding images of those elements. This is known as the *order-preserving* property.

Definition 3.1. A function $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ is *strictly-increasing* if for all $x, y \in \mathbb{R}^+ \cup \{0\}$, $f(x) < f(y)$ whenever $x < y$.

Throughout this thesis, the variables x, y, w, z are typically used to represent real numbers. There is no adherence to the function convention that y must represent a value in the range of a function and x must represent a value in the domain of a function (i.e. $y = f(x)$). Indeed, this function convention is avoided in this thesis as variables are nearly exclusively defined in the domain of a function.

Example 3.2. Consider the function $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ defined by $f(s) = \sqrt{s}$ for all $s \in \mathbb{R}^+ \cup \{0\}$ (shown in Figure 3.1). Consider $x, y \in \mathbb{R}^+ \cup \{0\}$ such that $x < y$. Then, $\sqrt{x} < \sqrt{y}$ and f is strictly-increasing.

3.2 Convex functions

In this thesis we are especially interested in convex functions. Informally, a function f is convex if the height of each point on the line segment between any two points on f is greater than or equal to the value of f at that point.

Definition 3.3. A function $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ is *convex* if, for all $x, y \in \mathbb{R}^+ \cup \{0\}$ and for all t where $0 \leq t \leq 1$, the inequity

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y)$$

holds.

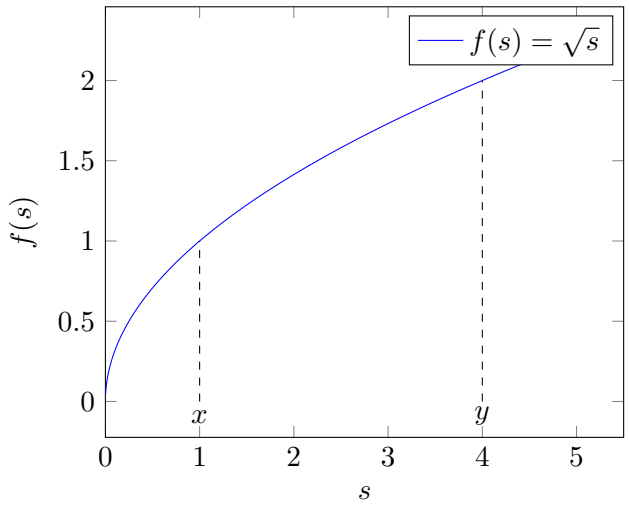


Figure 3.1: Plot of the function $f(s) = \sqrt{s}$ (in blue) and the real numbers x, y. The function f is strictly-increasing.

Example 3.4. Consider the function $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ defined by $f(s) = s^2$ for all $s \in \mathbb{R}^+ \cup \{0\}$ (depicted in Figure 3.2) and some $x, y \in \mathbb{R}^+ \cup \{0\}$. Let $t \in \mathbb{R}^+ \cup \{0\}$ with $0 \leq t \leq 1$. It can be shown that

$$f((1-t)x + t(y)) = (1-t)x^2 + ty^2 - (t - t^2)(x - y)^2.$$

Therefore, as $t - t^2 \geq 0$, the inequality

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y)$$

holds and f is convex. A visualisation of the convexity of f is shown in Figure 3.2.

The next result gives alternative but equivalent definitions of convexity; f is convex if it has non-decreasing gradient.

Proposition 3.5. *The following statements about a function $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ are equivalent:*

1. *The function f is convex.*
2. *For all $x, y \in \mathbb{R}^+ \cup \{0\}$ such that $x < y$, the inequality*
$$\frac{f(s) - f(x)}{s - x} \leq \frac{f(y) - f(x)}{y - x}$$
holds for all $s \in \mathbb{R}^+ \cup \{0\}$ where $x < s < y$.

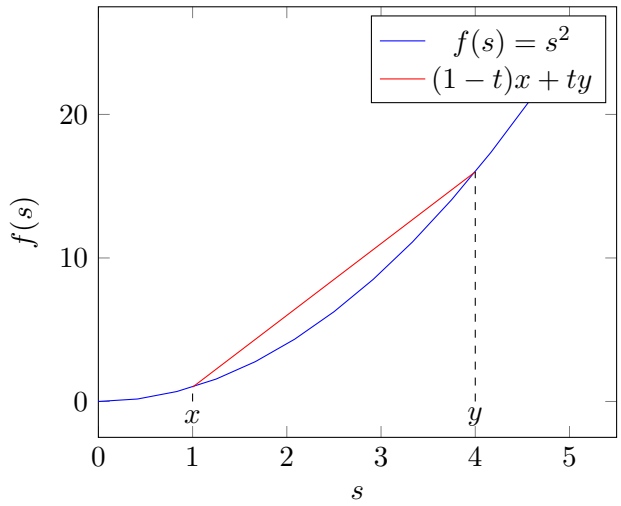


Figure 3.2: Plot of the convex function $f(s) = s^2$, real numbers x, y and $(1-t)x + ty$ for some t where $0 \leq t \leq 1$. The blue line represents the function f. The red line represents the segment from $(x, f(x))$ to $(y, f(y))$. Each point on the red line segment is directly above its corresponding point on f.

3. *For all* $x, y \in \mathbb{R}^+ \cup \{0\}$ *such that* $x < y$*, the inequality*

$$\frac{f(s) - f(x)}{s - x} \leq \frac{f(y) - f(s)}{y - s}$$

holds for all $s \in \mathbb{R}^+ \cup \{0\}$ *where* $x < s < y$.

Proof. Assume f is convex (Statement 1 holds). Consider $x, y \in \mathbb{R}^+ \cup \{0\}$ such that $x < y$. As f is convex

$$\begin{aligned} & f((1-t)x + ty) \leq (1-t)f(x) + tf(y) \\ \Rightarrow \quad & f(x + t(y-x)) \leq f(x) + t(f(y) - f(x)) \end{aligned}$$

for all t where $0 \leq t \leq 1$. Consider the substitution

$$t = \frac{s - x}{y - x}$$

for s where $x < s < y$. For all s in this range $0 < t < 1$ so

$$\begin{aligned} & f(x + \frac{s-x}{y-x}(y-x)) \leq f(x) + \frac{s-x}{y-x}(f(y) - f(x)) \\ \Rightarrow \quad & f(s) \leq f(x) + \frac{(s-x)(f(y) - f(x))}{y-x} \\ \Rightarrow \quad & \frac{f(s) - f(x)}{s-x} \leq \frac{f(y) - f(x)}{y-x} \end{aligned}$$

for all s where $x < s < y$. Further more,

$$\begin{aligned}
& \frac{f(s)-f(x)}{s-x} \leq \frac{f(y)-f(x)}{y-x} \\
\Leftrightarrow \quad & (f(s)-f(x))(y-x) \leq (f(y)-f(x))(s-x) \\
\Leftrightarrow \quad & yf(s)-xf(s)-yf(x)+xf(x) \leq sf(y)-xf(y)-sf(x)+xf(x) \\
\Leftrightarrow \quad & yf(s)-yf(x)+sf(x) \leq sf(y)-xf(y)+xf(s) \\
\Leftrightarrow \quad & yf(s)-yf(x)+sf(x)-sf(s) \leq sf(y)-xf(y)+xf(s)-sf(s) \\
\Leftrightarrow \quad & y(f(s)-f(x))-s(f(s)-f(x)) \leq s(f(y)-f(s))-x(f(y)-f(s)) \\
\Leftrightarrow \quad & (f(s)-f(x))(y-s) \leq (f(y)-f(s))(s-x) \\
\Leftrightarrow \quad & \frac{f(s)-f(x)}{s-x} \leq \frac{f(y)-f(s)}{y-s}
\end{aligned}$$

for all s where $x < s < y$. Hence Statement 1 implies Statement 2 and Statement 2 is equivalent to Statement 3. To complete the proof it is sufficient to show Statement 2 implies Statement 1.

Assume Statement 2 holds. Consider $x, y \in \mathbb{R}^+ \cup \{0\}$ and the three cases $x = y$, $x < y$ and $y < x$. If $x = y$ then

$$\begin{aligned}
f((1-t)x+ty) &= f(x-tx+tx) \\
&= f(x) \\
&= f(x)-tf(x)+tf(x) \\
&= (1-t)f(x)+tf(x) \\
&= (1-t)f(x)+tf(y)
\end{aligned}$$

for all t where $0 \leq t \leq 1$ (Statement 1 holds).

If $x < y$, consider the substitution $s = x + t(y-x)$ for t where $0 < t < 1$ in Statement 2. For t in this range $x < s < y$ so by Statement 2

$$
\begin{aligned}
& \frac{f(x+t(y-x))-f(x)}{x+t(y-x)-x} \leq \frac{f(y)-f(x)}{y-x} \\
\Rightarrow \quad & f(x+t(y-x))-f(x) \leq \frac{f(y)-f(x)}{y-x} t(y-x) \\
\Rightarrow \quad & f(x+t(y-x))-f(x) \leq t(f(y)-f(x)) \\
\Rightarrow \quad & f(ty-tx+x) \leq tf(y)-tf(x)+f(x) \\
\Rightarrow \quad & f((1-t)x+ty) \leq (1-t)f(x)+tf(y)
\end{aligned}
$$

for all t where $0 < t < 1$. When $t = 0$ or $t = 1$ convexity holds trivially as $f(x) = f(x)$ and $f(y) = f(y)$. Furthermore, if $y < x$, then the same argument can be applied with the substitution $t' = 1 - t$ in Statement 1. Hence Statement 2 implies Statement 1. □

Example 3.6. Consider the function $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ such that

$$
f(t) = \begin{cases} t, & \text{if } t < z \\ z + (t-z)\frac{w-x}{y-z}, & \text{if } t \geq z \end{cases}
$$

with $t \in \mathbb{R}^+ \cup \{0\}$. Suppose it were desirable to show f is strictly-increasing and convex under the assumptions $x \leq y$, $w \leq z$ and $0 < y - z \leq w - x$.

To establish f as strictly-increasing assume $t_0 < t_1$ for some $t_0, t_1 \in \mathbb{R}^+ \cup \{0\}$. If $t_1 \leq z$, then $f(t_0) = t_0 < t_1 = f(t_1)$. If $t_0 \geq z$, then

$$
\begin{aligned}
& t_0 < t_1 \\
\Rightarrow \quad & t_0 - z < t_1 - z \\
\Rightarrow \quad & (t_0 - z)\frac{w-x}{y-z} < (t_1 - z)\frac{w-x}{y-z} \\
\Rightarrow \quad & z + (t_0 - z)\frac{w-x}{y-z} < z + (t_1 - z)\frac{w-x}{y-z} \\
\Rightarrow \quad & f(t_0) < f(t_1).
\end{aligned}
$$

The fraction $\frac{w-x}{y-z}$ is defined and positive due to the premise $0 < y-z \leq w-x$. For the remaining case assume $t_0 < z$ and $t_1 > z$. Then,

$$
f(t_0) = t_0 < z < z + (t_1 - z)\frac{w-x}{y-z} = f(t_1),
$$

as $(t_1 - z)$ and $\frac{w-x}{y-z}$ are both non-negative. In all cases $f(t_0) < f(t_1)$ and so f is strictly-increasing.

To establish convexity, again assume $t_0 < t_1$ for some $t_0, t_1 \in \mathbb{R}^+ \cup \{0\}$. If $t_1 \leq z$ or $t_0 \geq z$, then convexity is trivial as both equations that define f are linear in t. For the remaining case assume $t_0 < z$ and $t_1 > z$, and consider $s \in \mathbb{R}^+ \cup \{0\}$ such that $t_0 < s < t_1$. If $s \leq z$, then

$$
\begin{aligned}
& 0 < y - z \leq w - x \\
\Rightarrow \quad & 1 \leq \frac{w-x}{y-z} \\
\Rightarrow \quad & z(\frac{w-x}{y-z} - 1) \leq t_1(\frac{w-x}{y-z} - 1) \\
\Rightarrow \quad & z\frac{w-x}{y-z} - z \leq t_1\frac{w-x}{y-z} - t_1 \\
\Rightarrow \quad & t_1 \leq t_1\frac{w-x}{y-z} - z\frac{w-x}{y-z} + z \\
\Rightarrow \quad & t_1 \leq z + (t_1 - z)\frac{w-x}{y-z} \\
\Rightarrow \quad & t_1 - s \leq z + (t_1 - z)\frac{w-x}{y-z} - s \\
\Rightarrow \quad & 1 \leq \frac{z + (t_1 - z)\frac{w-x}{y-z} - s}{t_1 - s} \\
\Rightarrow \quad & \frac{s - t_0}{s - t_0} \leq \frac{z + (t_1 - z)\frac{w-x}{y-z} - s}{t_1 - s} \\
\Rightarrow \quad & \frac{f(s) - f(t_0)}{s - t_0} \leq \frac{f(t_1) - f(s)}{t_1 - s}.
\end{aligned}
$$

Alternatively if $s > z$, then

$$\begin{aligned}
& 0 < y - z \leq w - x \\
\Rightarrow \quad & 1 \leq \frac{w-x}{y-z} \\
\Rightarrow \quad & t_0(\frac{w-x}{y-z} - 1) \leq z(\frac{w-x}{y-z} - 1) \\
\Rightarrow \quad & t_0\frac{w-x}{y-z} - t_0 \leq z\frac{w-x}{y-z} - z \\
\Rightarrow \quad & -z\frac{w-x}{y-z} - t_0 + z \leq -t_0\frac{w-x}{y-z} \\
\Rightarrow \quad & s\frac{w-x}{y-z} - z\frac{w-x}{y-z} - t_0 + z \leq s\frac{w-x}{y-z} - t_0\frac{w-x}{y-z} \\
\Rightarrow \quad & (s-z)\frac{w-x}{y-z} - t_0 + z \leq (s-t_0)\frac{w-x}{y-z} \\
\Rightarrow \quad & \frac{(s-z)\frac{w-x}{y-z} - t_0 + z}{s-t_0} \leq \frac{w-x}{y-z} \\
\Rightarrow \quad & \frac{(s-z)\frac{w-x}{y-z} - t_0 + z}{s-t_0} \leq \frac{(t_1-s)\frac{w-x}{y-z}}{t_1-s} \\
\Rightarrow \quad & \frac{(s-z)\frac{w-x}{y-z} - t_0 + z}{s-t_0} \leq \frac{(t_1-z)\frac{w-x}{y-z} - (s-z)\frac{w-x}{y-z}}{t_1-s} \\
\Rightarrow \quad & \frac{(s-z)\frac{w-x}{y-z} - t_0 + z}{s-t_0} \leq \frac{z + (t_1-z)\frac{w-x}{y-z} - z - (s-z)\frac{w-x}{y-z}}{t_1-s} \\
\Rightarrow \quad & \frac{f(s) - f(t_0)}{s-t_0} \leq \frac{f(t_1) - f(s)}{t_1-s}.
\end{aligned}$$

In each case convexity is shown by Proposition 3.

It is sometimes useful to warp convex functions with an operation that preserves convexity. This gives the inspiration for the next results, in which two such convexity-preserving operations are established.

Proposition 3.7. *Strict-increase and convexity is preserved under positive scaling. More formally, let $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ be a strictly-increasing convex function. If $g(x) = sf(x)$ for all $x \in \mathbb{R}^+ \cup \{0\}$ and some $s \in \mathbb{R}^+$, then g is a strictly-increasing convex function. Furthermore, if f is strictly-increasing, then g is a strictly-increasing function.*

Proof. Consider some $s \in \mathbb{R}^+$, some $t \in \mathbb{R}$ where $0 \leq t \leq 1$ and some $x, y \in \mathbb{R}^+ \cup \{0\}$. As s is positive and f is convex, the inequality

$$sf((1-t)x + ty) \leq s((1-t)f(x) + tf(y))$$

holds. Hence

$$g((1-t)x+ty) \leq (1-t)g(x)+tg(y),$$

and g is convex.

Suppose the restriction $x < y$ is applied. As s is positive and f is strictly-increasing $sf(x) < sf(y)$. Hence $g(x) < g(y)$, and g is strictly-increasing. □

Proposition 3.8. *The set of all strictly-increasing convex functions is closed under function addition. In particular, let* $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ *and* $g : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ *be strictly-increasing convex functions. If* $h(x) = f(x) + g(x)$ *for all* $x \in \mathbb{R}^+ \cup \{0\}$*, then* h *is a strictly-increasing convex function. Furthermore, if* f *and* g *are strictly-increasing functions, then* h *is a strictly-increasing function.*

Proof. Consider some $t \in \mathbb{R}$ where $0 \leq t \leq 1$ and some $x, y \in \mathbb{R}^+ \cup \{0\}$. As f and g are convex, the inequality

$$f((1-t)x+ty)+g((1-t)x+ty) \leq (1-t)f(x)+tf(y)+(1-t)g(x)+tg(y)$$

holds. Hence

$$h((1-t)x+ty) \leq (1-t)h(x)+th(y),$$

and h is convex.

Suppose the restriction $x < y$ is applied. As f and g are strictly-increasing, the inequality $f(x) + g(x) < f(y) + g(y)$ holds. Hence $h(x) < h(y)$, and h is strictly-increasing. □

Propositions 3.7 and 3.8 are useful for showing the existence of convex functions with certain properties by combining predefined convex functions. A simple example (Example 3.9) of this is given next. The example relies on the intermediate value theorem which states that a continuous function assumes every (intermediate) value between elements of the range. The intermediate value theorem is attributed to Bernard Bolzano in 1817 and the English translation of his original work was done by Russ (1980).

Example 3.9. Suppose the existence of a strictly-increasing convex function $h : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ with $h(0) = 0$, $h(1) = 1$ and $h(3) = 8$ was in

question. Consider the strictly-increasing convex functions $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ and $g : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ defined by $f(s) = s^2$ and $g(s) = s$ for all $s \in \mathbb{R}^+ \cup \{0\}$, respectively. The following properties hold: $f(0) = g(0) = 0$, $f(1) = g(1) = 1$, $f(3) = 9$, $g(3) = 3$. Both functions f and g have some of the properties that are desired of h. By Proposition 3.7 and Proposition 3.8 the function $h_t(s) = tg(s) + (1-t)f(s)$ is a strictly-increasing convex function for all $t \in \mathbb{R}^+$ such that $0 \leq t \leq 1$. Setting $s = 3$ and considering $h_t(3)$ as a function of t gives $h_0(3) = 3$ and $h_1(3) = 9$. Hence by the continuity of h (with respect to t) and the intermediate value theorem there exists a t_0 such that $h_{t_0}(3) = 8$. Hence there exists a strictly-increasing convex function $h = h_{t_0}$ with $h(0) = 0$, $h(1) = 1$ and $h(3) = 8$.

3.3 Dissimilarity maps

Definition 3.10. A *dissimilarity map* δ on a set X is a function $\delta : X \times X \to \mathbb{R}^+ \cup \{0\}$ that has the following properties for all $a, b \in X$:

1. $\delta(a, a) = 0$
2. $\delta(a, b) = \delta(b, a)$

In the context of functions, the first property is known as *reflexivity* and the second as *symmetry*. Let $X' \subset X$. Then $\delta|X'$ denotes the dissimilarity map δ restricted to the domain $X' \times X'$.

It is important to note that in many fields, such as algorithmic computer science, dissimilarity maps are known as distance matrices. Dissimilarity map is the term consistent with Semple and Steel (2003), the main notational reference for this thesis.

Example 3.11. Let the set $X = \{a, b, c, d, e, f\}$ and $\delta : X \times X \to \mathbb{R}^+ \cup \{0\}$

be the function

$$\delta = \begin{array}{c|cccccc} & a & b & c & d & e & f \\ \hline a & 0 & 1 & 2 & 3 & 4 & 0 \\ b & 1 & 0 & 2 & 0 & 3 & 0 \\ c & 2 & 2 & 0 & 3 & 2 & 0 \\ d & 3 & 0 & 3 & 0 & 4 & 0 \\ e & 4 & 3 & 2 & 4 & 0 & 5 \\ f & 0 & 0 & 0 & 0 & 5 & 0 \end{array}.$$

Each entry in the main diagonal is zero, so δ is reflexive. Similarly, δ is symmetric about the main diagonal, so δ is symmetric. Hence δ is a dissimilarity map.

3.4 Tree metrics

Definition 3.12. A *metric* δ on a set X is a dissimilarity map that has the following additional properties for all $a, b, c \in X$:

1. if $\delta(a, b) = 0$, then $a = b$
2. $\delta(a, c) \leq \delta(a, b) + \delta(b, c)$

The second property (Property 2) is known as the *triangle inequality*. If just the triangle inequality holds, then δ is a *pseudometric*. Alternatively, if just the first property holds, then δ is a *strong dissimilarity map*. If both properties hold and the triangle inequality is strengthened to $\delta(a, c) < \delta(a, b) + \delta(b, c)$ (for distinct $a, b, c \in X$), then δ is a *strict metric*.

Example 3.13. Let the set $X = \{a, b, c\}$ and $\delta_i : X \times X \to \mathbb{R}^+ \cup \{0\}$ for $i \in \{1, 2, 3, 4\}$ be the dissimilarity maps defined by:

$$\delta_1 = \begin{array}{c|ccc} & a & b & c \\ \hline a & 0 & 0 & 3 \\ b & 0 & 0 & 1 \\ c & 3 & 1 & 0 \end{array} \quad \delta_2 = \begin{array}{c|ccc} & a & b & c \\ \hline a & 0 & 0 & 3 \\ b & 0 & 0 & 3 \\ c & 3 & 3 & 0 \end{array} \quad \delta_3 = \begin{array}{c|ccc} & a & b & c \\ \hline a & 0 & 2 & 4 \\ b & 2 & 0 & 2 \\ c & 4 & 2 & 0 \end{array} \quad \delta_4 = \begin{array}{c|ccc} & a & b & c \\ \hline a & 0 & 2 & 3 \\ b & 2 & 0 & 4 \\ c & 3 & 4 & 0 \end{array}.$$

The dissimilarity map δ_1 has neither metric property. For the dissimilarity map δ_2 the triangle inequality holds but $\delta_2(a, b) = 0$, so δ_2 is a

pseudometric. Both δ_3 and δ_4 are metrics but only δ_4 is a strict metric as

$$\delta_3(a, c) = 4 = 2 + 2 = \delta_3(a, b) + \delta_3(b, c).$$

In this thesis the empty sum convention of $\sum_{\emptyset} = 0$ is used. That is, given an empty set of elements, a real number summation over that set will sum to 0. This convention is relevant particularly to the next definition, where empty summations are not explicitly dealt with.

Definition 3.14. For a given tree T and weight function $w : E(T) \to \mathbb{R}^+$, the pseudometric $d_{(T,w)}$ on $V(T)$ is defined as follows:

$$d_{(T,w)}(u, v) = \sum_{e \in E(P_T(u,v))} w(e) \text{ for all } u, v \in V(T),$$

where $E(P_T(u, v))$ is the set of edges on the (unique) path $P_T(u, v)$ from u to v in T. Similarly, for an X-tree $\mathcal{T} = (T, \phi)$, the pseudometric $d_{(\mathcal{T},w)}$ on X is defined:

$$d_{(\mathcal{T},w)}(a, b) = d_{(T,w)}(\phi(a), \phi(b)) \text{ for all } a, b \in X.$$

A dissimilarity map δ on X is a *tree metric* if there exists an X-tree $\mathcal{T} = (T, \phi)$ with edge set $E(T)$ and a weight function $w : E(T) \to \mathbb{R}^+$ such that

$$\delta(a, b) = d_{(\mathcal{T},w)}(a, b) \text{ for all } a, b \in X.$$

In this case, the tree metric δ is also known as a $\mathcal{T}$-metric. Furthermore, the pair $(\mathcal{T}, w)$ is a *tree metric representation* of δ and $\mathcal{T}$ *supports* δ.

Not all tree metrics are metrics. Some tree metrics $\delta : X \times X \to \mathbb{R}^+ \cup \{0\}$ have $\delta(a, b) = 0$ for some distinct $a, b \in X$. Hence a tree metric is only guaranteed to be a pseudometric.

Example 3.15. Let the set $X = \{a, b, c, d, e, f\}$. Consider the weighted X-tree $(\mathcal{T}, w)$ shown in Figure 3.3. The pseudometric $d_{(\mathcal{T},w)} : X \times X \to \mathbb{R}^+ \cup \{0\}$ is given by

$$d_{(\mathcal{T},w)} = \begin{array}{c|ccccccc} & a & b & c & d & e & f & g \\ \hline a & 0 & 7 & 8 & 6 & 6 & 11 & 9 \\ b & 7 & 0 & 5 & 3 & 3 & 8 & 6 \\ c & 8 & 5 & 0 & 4 & 4 & 9 & 7 \\ d & 6 & 3 & 4 & 0 & 0 & 5 & 3 \\ e & 6 & 3 & 4 & 0 & 0 & 5 & 3 \\ f & 11 & 8 & 9 & 5 & 5 & 0 & 8 \\ g & 9 & 6 & 7 & 3 & 3 & 8 & 0 \end{array}.$$

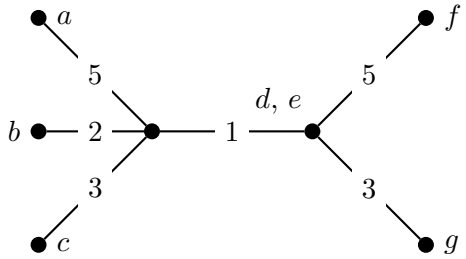


Figure 3.3: The weighted X-tree $(\mathcal{T}, w)$.

Theorem 3.16 allows us to define an X-tree and a weight function in terms of a tree metric. Conversely, it is trivial to check that an X-tree weight function pair $(\mathcal{T}, w)$ induces a unique tree metric $d_{(\mathcal{T},w)}$ on X. Because of the analogous relationship between tree metric representations and tree metrics we often make no effort to distinguish between the two. The proof of the uniqueness result (Theorem 3.16) is given in Semple and Steel (2003, p. 148).

Theorem 3.16. *Let δ be a tree metric on X. Then, there is (up to isomorphism) exactly one tree metric representation of δ.*

The following proposition exposes the conditions for a tree metric to have a phylogenetic X-tree support it (as opposed to a more general X-tree).

Proposition 3.17. *Let $d_{(\mathcal{T},w)}$ be a tree ($\mathcal{T}$-)metric on X. The X-tree $\mathcal{T}$ is a phylogenetic X-tree on X if and only if $d_{(\mathcal{T},w)}$ is a strict metric.*

Proof. Consider the case where $\mathcal{T} = (T, \phi)$ is a phylogenetic X-tree and a, b and c are distinct elements of X. As all edges have positive weights (under the weight function w) and there is at least one edge between the leaves $\phi(a)$ and $\phi(b)$, the inequality $d_{(\mathcal{T},w)}(a, b) > 0$ holds. Furthermore, trees have no cycles (and therefore unique paths between all vertices), hence $d_{(\mathcal{T},w)}$ is a metric. Let v be the (unique) internal vertex the paths $\phi(a)$ to $\phi(b)$, $\phi(b)$ to $\phi(c)$ and $\phi(a)$ to $\phi(c)$ have in common. As each of $\phi(a), \phi(b)$ and $\phi(c)$ are leaves, the distance from each to v is strictly positive. Hence, as paths between vertices are unique, $d_{(\mathcal{T},w)}$ is a strict metric.

Conversely, consider the case where the X-tree $\mathcal{T} = (T, \phi)$ is not a phylogenetic X-tree. Then, either there exists an internal vertex $v \in V$ of $T = (V, E)$ in the range of ϕ, or ϕ is not injective from X to the leaves of T (i.e. it assigns multiple labels to the same leaf in T). In the first case, the strict triangle inequality is broken and thus $d_{(\mathcal{T},w)}$ is not a strict metric. In the remaining case (where ϕ is not injective), there exist elements $a, b \in X$ such that $\phi(a) = \phi(b)$ and thus $d_{(\mathcal{T},w)}(a, b) = 0$. In either case $d_{(\mathcal{T},w)}$ is not a strict metric. □

Definition 3.18. Let δ be a dissimilarity map on X. For every four elements $a, b, c, d \in X$, if the two greatest (or greatest equal) sums from $\delta(a, b) + \delta(c, d)$, $\delta(a, c) + \delta(b, d)$ and $\delta(a, d) + \delta(b, c)$ are equal then δ satisfies the *four-point condition*. For distinct elements $a, b, c, d \in X$, the dissimilarity map $\delta|\{a, b, c, d\}$ has *type* $ab|cd$ if the sum $\delta(a, b) + \delta(c, d)$ is smaller than the two equal sums $\delta(a, c) + \delta(b, d)$ and $\delta(a, d) + \delta(b, c)$. If all three sums are equal $\delta|\{a, b, c, d\}$, then has type $abcd$.

The four-point condition has to apply to all subsets of X of size less than four also, as the elements from X need not be distinct. Therefore the four-point condition implies the triangle inequality. To see this consider some set X, a dissimilarity map δ and any three elements $a, b, c \in X$. By the four-point condition applied to a, b, b and c, the greatest (or greatest equal) two sums of $\delta(a, c) + \delta(b, b)$, $\delta(a, b) + \delta(b, c)$ and $\delta(a, b) + \delta(b, c)$ are equal. Hence, $\delta(a, c) \leq \delta(a, b) + \delta(b, c)$.

In brief, the four-point condition for three elements (with one repeated) is equivalent to the triangle inequality. Similarly, the triangle inequality and

the four-point condition holding for every set of four distinct vertices implies the four-point condition. A consequence of this is that the dissimilarity maps that satisfy the four-point condition are at least pseudometrics.

Example 3.19. Let δ be a dissimilarity map on $X = \{a, b, c, d\}$ given by

$$\delta = \begin{array}{c|cccc} & a & b & c & d \\ \hline a & 0 & 9 & 12 & 15 \\ b & 9 & 0 & 17 & 18 \\ c & 12 & 17 & 0 & 23 \\ d & 15 & 18 & 23 & 0 \end{array}.$$

The triangle inequality holds for δ. The inequality

$$\delta(a, c) + \delta(b, d) = 30 < 32 = \delta(a, d) + \delta(b, c) = \delta(a, b) + \delta(c, d)$$

also holds, so the four-point condition holds for every four distinct elements of X. Hence the four-point condition holds for δ, and δ has type $ac|bd$.

Proposition 3.20. *Let δ be a strict metric on $X = \{a, b, c, d\}$. The metric δ is a $ab|cd$-metric if and only if δ has type $ab|cd$. Furthermore, δ is a $abcd$-metric if and only if δ has type $abcd$.*

Proof. Suppose the phylogenetic X-tree $\mathcal{T} = ab|cd$ supports δ using some weight function w and $\mathcal{T} = (T, \phi)$. Let i' denote $\phi(i)$ for all $i \in X$. Let u and v be the internal vertices of $\mathcal{T}$ adjacent to a' (and b') and adjacent to c' (and d') respectively. Then

$$\delta(a, c)+\delta(b, d) = w(\{a', u\})+w(\{b', u\})+2w(\{u, v\})+w(\{c', v\})+w(\{d', v\}),$$

$$\delta(a, d)+\delta(b, c) = w(\{a', u\})+w(\{b', u\})+2w(\{u, v\})+w(\{c', v\})+w(\{d', v\})$$

and

$$\delta(a, b) + \delta(c, d) = w(\{a', u\}) + w(\{b', u\}) + w(\{c', v\}) + w(\{d', v\}).$$

As weight functions are strictly positive and the first two sums are equal, δ has type $ab|cd$. A simplified version of this argument can be used for $\mathcal{T} = abcd$ and δ having type $abcd$.

Conversely, suppose δ has type $ab|cd$. Consider the tree T given in Figure 3.4 (a binary tree on $\{v_0, v_1, v_2, v_3, v_4, v_5\}$ with exactly two internal vertices v_4 and v_5 and with v_0 and v_1 adjacent to v_4), the weight function w on $\{\{v_0, v_4\}, \{v_1, v_4\}, \{v_2, v_5\}, \{v_3, v_5\}, \{v_4, v_5\}\}$ defined by

$E(T)$	$w(E(T))$
$\{v_0, v_4\}$	$\frac{\delta(a,c)+\delta(a,b)-\delta(b,c)}{2}$
$\{v_1, v_4\}$	$\frac{\delta(b,c)+\delta(a,b)-\delta(a,c)}{2}$
$\{v_2, v_5\}$	$\frac{\delta(a,c)+\delta(c,d)-\delta(a,d)}{2}$
$\{v_3, v_5\}$	$\frac{\delta(a,d)+\delta(c,d)-\delta(a,c)}{2}$
$\{v_4, v_5\}$	$\frac{\delta(a,c)+\delta(b,d)-(\delta(a,b)+\delta(c,d))}{2}$

,

and the labelling function ϕ given by

$$\phi = \begin{array}{c|cccc} X & a & b & c & d \\ \hline \phi(X) & v_0 & v_1 & v_2 & v_3 \end{array}.$$

But $ab|cd = (T, \phi)$ and w has $d_{(T,w)}(\phi(\alpha), \phi(\beta)) = \delta(\alpha, \beta)$ for all $\alpha, \beta \in X$, hence δ is an $ab|cd$-metric. The same argument without the edge $\{v_4, v_5\}$ yields the result for δ with type $abcd$.

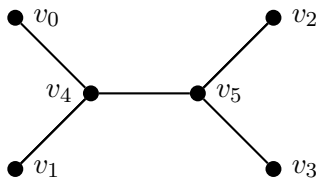


Figure 3.4: The tree T used in the proof of Proposition 3.20.

□

The four-point condition allows a widely used characterization of tree metrics attributed to Zaretskii (1965). A modern proof of the same result is contained in Semple and Steel (2003, p. 152). This is the foundation and focus for most reconstruction methods from dissimilarity information.

Theorem 3.21. *A dissimilarity map δ is a tree metric if and only if it satisfies the four-point condition.*

In the following example, the process for reconstructing a weighted X-tree from a dissimilarity map that satisfies the four-point condition is explained. The same general construction can be used in the proof of Theorem 3.21.

Example 3.22. Consider the dissimilarity map δ on $X = \{a, b, c, d\}$ given by

$$\delta = \begin{array}{c|cccc} & a & b & c & d \\ \hline a & 0 & 9 & 12 & 15 \\ b & 9 & 0 & 17 & 18 \\ c & 12 & 17 & 0 & 23 \\ d & 15 & 18 & 23 & 0 \end{array}.$$

The four-point condition for distinct vertices holds as $\delta(a, b) + \delta(c, d) = 32$, $\delta(a, c) + \delta(b, d) = 30$ and $\delta(a, d) + \delta(b, c) = 32$. It also holds for repeated vertices as the triangle inequality holds. Hence, Theorem 3.21 can be applied and δ is a tree ($\mathcal{T}$-)metric for some X-tree $\mathcal{T}$. Furthermore, Proposition 3.17 shows $\mathcal{T}$ is a phylogenetic X-tree as δ is a strict metric.

The question now becomes which phylogenetic X-tree is $\mathcal{T}$? Theorem 3.16 and Proposition 2.23 narrow the possibilities to one of the phylogenetic X-trees $ab|cd$, $ac|bd$, $ad|bc$ or $abcd$. Proposition 3.20 confirms that $\mathcal{T}$ is isomorphic to $ac|bd$ as δ has type $ac|bd$.

3.5 Connecting functions and trees

Definition 3.23. A dissimilarity map δ_1 on X is *order equivalent* to a dissimilarity map δ_2 on X if there is a strictly-increasing function $f : \mathbb{R} \to \mathbb{R}$ such that $\delta_1(a, b) = f(\delta_2(a, b))$ for all $a, b \in X$. Order equivalence is denoted $\delta_1 \overset{o.e.}{\sim} \delta_2$. The ordinal assumption is that order equivalence holds between a given (directed) pair of dissimilarity maps.

Definition 3.24. A dissimilarity map δ_1 on X is *convex related* to a dissimilarity map δ_2 on X if there is a strictly-increasing convex function $f : \mathbb{R} \to \mathbb{R}$ such that $\delta_1(a, b) = f(\delta_2(a, b))$ for all $a, b \in X$. Such a convex relation is denoted $\delta_1 \overset{c.r.}{\to} \delta_2$. The convex assumption is that the convex relation holds between a given (directed) pair of dissimilarity maps.

In the previous two definitions statements of the form $\delta_1(a,b) = f(\delta_2(a,b))$ for all $a,b \in X$ are used. In this thesis the abbreviated form $\delta_1 = f(\delta_2)$ is often used with the same meaning.

Example 3.25. Let the dissimilarity maps δ_1 and δ_2 on X be defined:

$$\delta_1 = \begin{array}{c|ccccc} & a & b & c & d & e \\ \hline a & 0 & 1 & 9 & 3 & 6 \\ b & 1 & 0 & 7 & 0 & 5 \\ c & 9 & 7 & 0 & 4 & 2 \\ d & 3 & 0 & 4 & 0 & 8 \\ e & 6 & 5 & 2 & 8 & 0 \end{array} \quad \text{and} \quad \delta_2 = \begin{array}{c|ccccc} & a & b & c & d & e \\ \hline a & 0 & 2 & 14 & 4 & 11 \\ b & 2 & 0 & 12 & 0 & 10 \\ c & 14 & 12 & 0 & 6 & 3 \\ d & 4 & 0 & 6 & 0 & 13 \\ e & 11 & 10 & 3 & 13 & 0 \end{array}.$$

Let f be any function such that $\delta_1(a,b) = f(\delta_2(a,b))$ for all $a,b \in X$. The function f has decreasing gradient as $f(4) = 6$, $f(5) = 10$ and $f(6) = 11$. Therefore f can't be convex (by Proposition 3.5) and δ_1 is not convex related to δ_2. If f is linearly interpolated (points joined by straight-line segments) over successive $(\delta_1(a,b), \delta_2(a,b))$ points as depicted in Figure 3.5, then f is a strictly-increasing function from δ_1 to δ_2. Hence $\delta_1 \overset{o.e.}{\sim} \delta_2$.

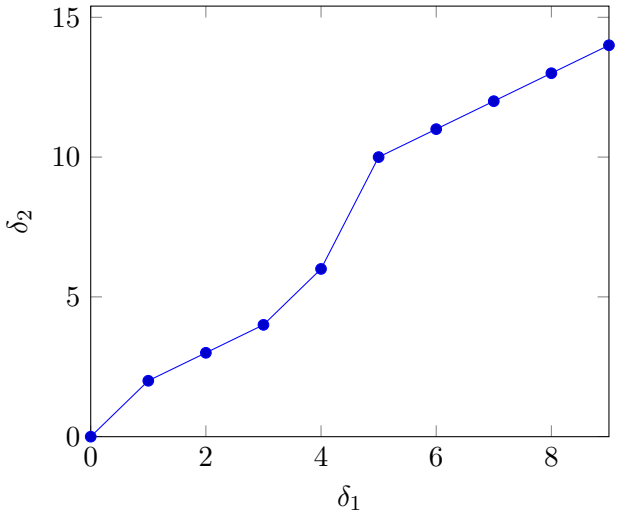


Figure 3.5: The order-preserving function f from δ_1 to δ_2 used in Example 3.25.

Filling the 'gaps' between points using a function with straight line segments (as done in Example 3.25) is known *linear interpolation*. It can be

shown that there is an order-preserving function fitting a sequence of points if and only if the linear interpolation is order-preserving. Similarly, using the mean value theorem it is possible to prove there is a (continuous) convex function fitting a sequence of points if and only if the linear interpolation is convex. One way to prove the backwards implication of this result is to assume the linear interpolation between the points is not convex, then, apply the mean value theorem on each line segment to show that no (continuous) function can simultaneously have strictly-increasing gradient and fit the interpolation points. Therefore, the linear interpolation of ordered pairs of corresponding entries from the two dissimilarity maps is the only function that needs to be considered when proving the existence (or non-existence) of an order-preserving or convex function. More briefly, examining many functions between dissimilarity maps is redundant as the linear interpolation gives all necessary details. In this thesis unless explicitly defined, a particular function between two dissimilarity maps is assumed to be the linear interpolation of ordered pairs of corresponding entries from the two dissimilarity maps.

A linear interpolation between points is considered to maintain the gradient between the two closest distinct points for all unknown parts of the function's domain. It is important to note that for a linear interpolation to be well-defined the domain must be given in non-decreasing order. Furthermore, no distinct points can have the same domain value (e.g. the linear interpolation between $(1, 2)$ and $(1, 3)$ is not a well-defined function). Lastly, at least two distinct domain values must be given for the linear interpolation to be a well-defined function.

Definition 3.26. Let δ be a dissimilarity map on X. If there is an X-tree $\mathcal{T}$ and a weight function w such that $\delta \overset{o.e.}{\sim} d_{(\mathcal{T},w)}$, then δ *fits* $\mathcal{T}$ under order equivalence. Similarly if $\delta \overset{c.r.}{\to} d_{(\mathcal{T},w)}$, then δ *fits* $\mathcal{T}$ under convex relation. If $\mathcal{T}$ is the only phylogenetic X-tree δ fits, then δ *defines* $\mathcal{T}$ (under the respective relation). Let $\mathbf{T}_\delta^{o.e.}$ and $\mathbf{T}_\delta^{c.r.}$ be the set of all phylogenetic X-trees fitted by δ under order equivalence and convex relation respectively.

Example 3.27. Consider the two dissimilarity maps δ_1 and δ_2 on X

$$\delta_1 = \begin{array}{c|cccc} & a & b & c & d \\ \hline a & 0 & 6 & 10.5 & 14 \\ b & 6 & 0 & 12 & 15 \\ c & 10.5 & 12 & 0 & 9 \\ d & 14 & 15 & 9 & 0 \end{array} \quad \text{and} \quad \delta_2 = \begin{array}{c|cccc} & a & b & c & d \\ \hline a & 0 & 6 & 12 & 18 \\ b & 6 & 0 & 14 & 20 \\ c & 12 & 14 & 0 & 10 \\ d & 18 & 20 & 10 & 0 \end{array}.$$

Using Theorem 3.21 and Proposition 3.20, the dissimilarity map δ_2 can be shown to be a $\mathcal{T}$-metric for the phylogenetic X-tree $\mathcal{T} = ab|cd$ (with weight function depicted in Figure 3.6) as the four-point condition holds and δ_2 has type $ab|cd$. But the function $f : \delta_1 \to \delta_2$ is convex as the gradient of each line segment forms the non-decreasing sequence $(1, 1.5, 1.5, 1.5, 2, 2)$. Hence, δ_1 fits $ab|cd$ under the convex relation and $ab|cd \in \mathbf{T}^{c.r.}_{\delta_1}$. The function f is depicted in Figure 3.7.

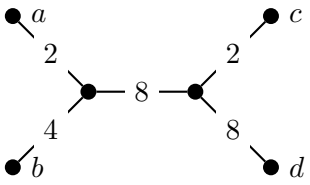


Figure 3.6: The phylogenetic X-tree $ab|cd$ with corresponding weight function.

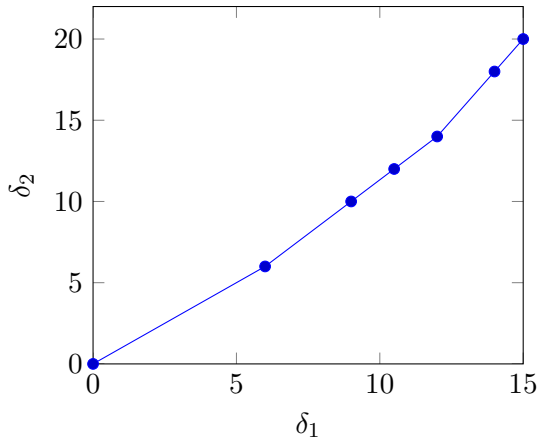


Figure 3.7: The strictly-increasing convex function f from δ_1 to δ_2 used in Example 3.27.

What is the connection between $\mathbf{T}^{c.r.}_{\delta}$ and $\mathbf{T}^{o.e.}_{\delta}$ for some dissimilarity map δ? Proposition 3.28 is based on the trivial observation that strictly-increasing convex functions are also strictly-increasing functions.

Proposition 3.28. *If δ is a dissimilarity map on X then $\mathbf{T}^{c.r.}_{\delta} \subseteq \mathbf{T}^{o.e.}_{\delta}$.*

Proof. If $\mathbf{T}^{c.r.}_{\delta}$ is empty the result yields. Assume $\mathbf{T}^{c.r.}_{\delta}$ is non-empty. Consider a phylogenetic X-tree $\mathcal{T}$ in $\mathbf{T}^{c.r.}_{\delta}$. There exists a strictly-increasing

convex function $f : \mathbb{R} \to \mathbb{R}$ that brings δ to a $\mathcal{T}$-metric. As f is strictly-increasing $\mathcal{T} \in \mathbf{T}_\delta^{o.e.}$. Hence $\mathbf{T}_\delta^{c.r.} \subseteq \mathbf{T}_\delta^{o.e.}$. □

Corollary 3.29. *Let δ be a dissimilarity map on X such that δ defines a phylogenetic X-tree $\mathcal{T}$ under ordinal equivalence then either δ defines a phylogenetic X-tree $\mathcal{T}$ under convex relation or $\mathbf{T}_\delta^{c.r.}$ is empty.*

Proof. If δ defines an X-tree $\mathcal{T}$ under ordinal equivalence then $\mathbf{T}_\delta^{o.e.} = \{\mathcal{T}\}$. Hence by Proposition 3.28 $\mathbf{T}_\delta^{c.r.} = \emptyset$ or $\mathbf{T}_\delta^{c.r.} = \{\mathcal{T}\}$. □

Despite the previous two results, there is no guarantee that $\mathbf{T}_\delta^{c.r.}$ can be a proper subset of $\mathbf{T}_\delta^{o.e.}$. Example 3.30 gives an (unproven) case where this situation occurs. The tools needed to prove that $\mathbf{T}_\delta^{c.r.}$ is a proper subset of $\mathbf{T}_\delta^{o.e.}$ in this case are developed over the next two chapters.

Example 3.30. Let $X = \{a, b, c, d\}$ and

$\delta =$

	a	b	c	d
a	0	9	12	15
b	9	0	17	18
c	12	17	0	22
d	15	18	22	0

be a dissimilarity map on X. Consider the four weighted phylogenetic X-trees $(\mathcal{T}_1, w_1)$, $(\mathcal{T}_2, w_2)$, $(\mathcal{T}_3, w_3)$ and $(\mathcal{T}_4, w_4)$ with tree metrics

$d_{(\mathcal{T}_1,w_1)} =$

	a	b	c	d
a	0	4	9	10
b	4	0	11	12
c	9	11	0	13
d	10	12	13	0

, $d_{(\mathcal{T}_2,w_2)} =$

	a	b	c	d
a	0	9	12	15
b	9	0	17	18
c	12	17	0	23
d	15	18	23	0

,

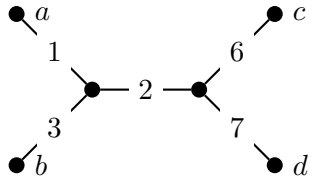
$d_{(\mathcal{T}_3,w_3)} =$

	a	b	c	d
a	0	6	8	9
b	6	0	10	13
c	8	10	0	15
d	9	13	15	0

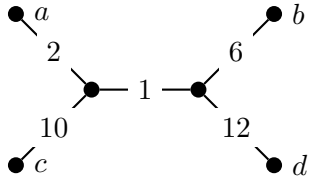
and $d_{(\mathcal{T}_4,w_4)} =$

	a	b	c	d
a	0	4	5	6
b	4	0	7	8
c	5	7	0	9
d	6	8	9	0

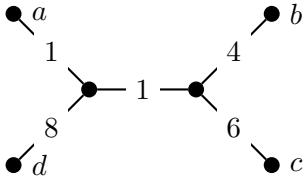
as depicted in Figure 3.8. These four trees represent every possible phylogenetic X-tree. Each function f_i from δ to $d_{(\mathcal{T}_i,w_i)}$, with $i \in \{1,2,3,4\}$, is shown in Figure 3.9. Each function is strictly-increasing and thus $\{\mathcal{T}_1,\mathcal{T}_2,\mathcal{T}_3,\mathcal{T}_4\} \subseteq \mathbf{T}_\delta^{o.e.}$. As there are no other phylogenetic X-trees with $|X| = 4$ (Proposition 2.23), the set $\{\mathcal{T}_1,\mathcal{T}_2,\mathcal{T}_3,\mathcal{T}_4\} = \mathbf{T}_\delta^{o.e.}$. Only one of the functions (Figure 3.9b) is convex. Hence $\mathcal{T}_2 \in \mathbf{T}_\delta^{c.r.}$. In later chapters, tools are developed that show $\mathbf{T}_\delta^{c.r.} = \{\mathcal{T}_2\}$ and that $\mathbf{T}_\delta^{c.r.}$ in this case is a proper subset of $\mathbf{T}_\delta^{o.e.}$.



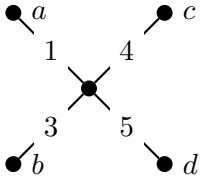
(a) The weighted binary phylogenetic X-tree $(\mathcal{T}_1, w_1)$ with $\mathcal{T}_1$ isomorphic to $ab|cd$.



(b) The weighted binary phylogenetic X-tree $(\mathcal{T}_2, w_2)$ with $\mathcal{T}_2$ isomorphic to $ac|bd$.

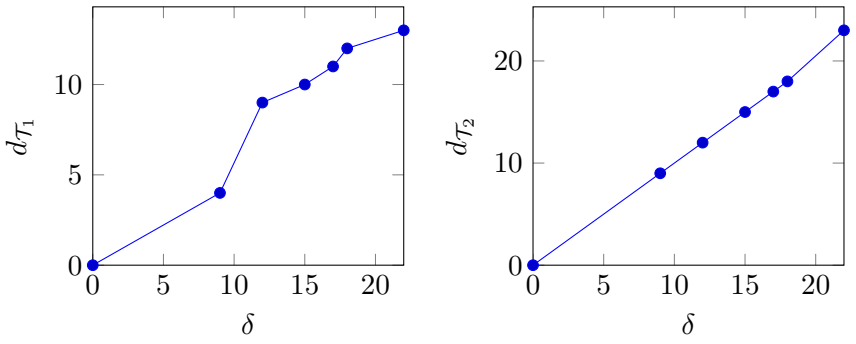


(c) The weighted binary phylogenetic X-tree $(\mathcal{T}_3, w_3)$ with $\mathcal{T}_3$ isomorphic to $ad|bc$.



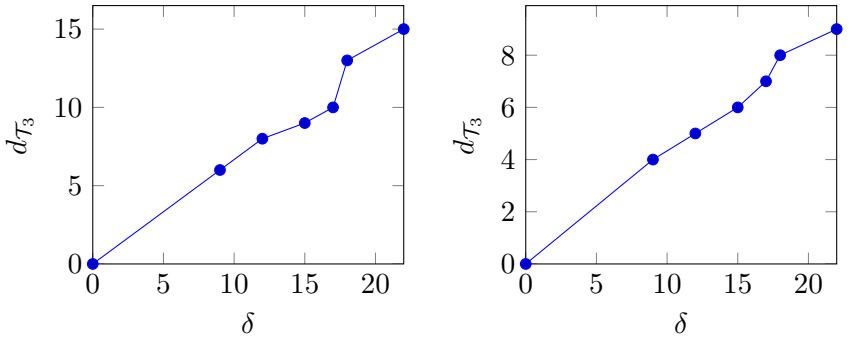
(d) The weighted (star) phylogenetic X-tree $(\mathcal{T}_4, w_4)$ with $\mathcal{T}_4$ isomorphic to $abcd$.

Figure 3.8: Four distinct phylogenetic X-trees with weight functions.



(a) Function from δ to $d_{(\mathcal{T}_1,w_1)}$.

(b) Function from δ to $d_{(\mathcal{T}_2,w_2)}$.



(c) Function from δ to $d_{(\mathcal{T}_3,w_3)}$.

(d) Function from δ to $d_{(\mathcal{T}_4,w_4)}$.

Figure 3.9: Four order-preserving functions from the dissimilarity map δ to a distinct tree metric.

Chapter 4

Interval relations

To help prove the correctness of the ordinal quartet method (Kearney, 1998), Kearney used an order interval relation. This relation helped show when it was possible to have a strictly-increasing function from one dissimilarity map to another. This chapter introduces the same interval relation (Section 4.1) and also introduces a new convex interval relation. In addition, a number of properties and examples involving these relations are presented (Sections 4.3.1 and 4.3.2). The chapter concludes with the introduction of 'clue' functions (Section 4.4), which have properties to aid in the classification of dissimilarity maps under the convex assumption. Essentially, this chapter serves to develop the tools necessary to prove the classification results in Chapter 5.

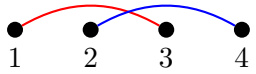
We begin by formalising the notion of an interval.

Definition 4.1. Given $x, y \in \mathbb{R}^+ \cup \{0\}$ with $x \leq y$ the *interval* between x and y is denoted $[x, y]$. In the case where $x = y$, the interval $[x, y]$ is a *trivial* interval.

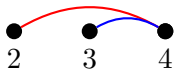
4.1 Order interval relation

Definition 4.2. The *order interval relation* $<_o$ on the set of all intervals is the relation such that $[x, y] <_o [w, z]$ if and only if $x < w$ and $y \leq z$, or $x \leq w$ and $y < z$, where $x, y, w, z \in \mathbb{R}^+ \cup \{0\}$.

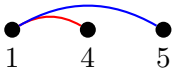
Example 4.3. The statements $[1,3] <_o [2,4]$, $[2,4] <_o [3,4]$ and $[1,4] <_o [1,5]$ all hold (as visualised in Figure 4.1). None of the statements $[1,3] <_o [1,3]$, $[2,3] <_o [1,5]$ or $[1,5] <_o [2,3]$ (visualised in Figure 4.2) hold.



(a) The intervals $[1,3]$ (in red) and $[2,4]$ (in blue).

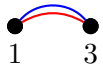


(b) The intervals $[2,4]$ (in red) and $[3,4]$ (in blue).

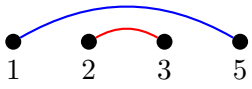


(c) The intervals $[1,4]$ (in red) and $[1,3]$ (in blue).

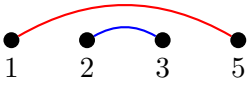
Figure 4.1: Pairs of intervals in which the red interval I_{red} and the blue interval I_{blue} satisfy $I_{red} <_o I_{blue}$.



(a) The intervals $[1,3]$ (in red) and $[1,3]$ (in blue).



(b) The intervals $[2,3]$ (in red) and $[1,5]$ (in blue).



(c) The intervals $[1,5]$ (in red) and $[2,3]$ (in blue).

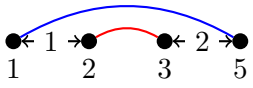
Figure 4.2: Pairs of intervals in which the red interval I_{red} and the blue interval I_{blue} satisfy neither $I_{red} <_o I_{blue}$ nor $I_{blue} <_o I_{red}$.

4.2 Convex interval relation

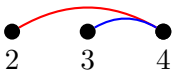
Definition 4.4. The *convex interval relation* $<_c$ on the set of all intervals is the relation such that $[x,y] <_c [w,z]$ if and only if $x < w$ and $y \leq z$, $x \leq w$ and $y < z$, or $0 < x - w < z - y$, where $x, y, w, z \in \mathbb{R}^+ \cup \{0\}$.

The last possibility $0 < x - w < z - y$ is the only difference between the order interval relation ($<_o$) and the convex interval relation ($<_c$). As a consequence the order of the elements in the intervals is not enough to determine whether the relation holds; comparative distances between interval start and end points need to be taken into account. Further in this chapter (Proposition 4.14), it is shown that this extra condition is the boundary condition on the existence of a convex function $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ such that $f(x) + f(y) \geq f(z) + f(w)$ and with $f(0) = 0$.

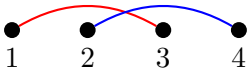
Example 4.5. The relations $[1,3] <_c [2,4]$, $[2,4] <_c [3,4]$ and $[2,3] <_c [1,5]$ all hold (even though $[2,3] <_o [1,5]$ does not hold). The intervals are visualised in Figure 4.3.



(c) The intervals $[2,3]$ (in red) and $[1,5]$ (in blue). The distance between 1 and 2 is smaller then the distance from 3 to 5 so $[2,3] <_c [1,5]$



(b) The intervals $[2,4]$ (in red) and $[3,4]$ (in blue).



(a) The intervals $[1,3]$ (in red) and $[2,4]$ (in blue).

Figure 4.3: Pairs of intervals in which the red interval I_{red} and the blue interval I_{blue} satisfy $I_{red} <_c I_{blue}$.

4.3 Interval relation properties

Definition 4.6. A *strict partial order* on a set A is any irreflexive, anti-symmetric and transitive relation $\mathcal{R}$ on A. If $x\mathcal{R}y$ or $y\mathcal{R}x$, then x and y are *$\mathcal{R}$-comparable* by the strict partial order $\mathcal{R}$. Conversely if x and y are not comparable (by the strict partial order $\mathcal{R}$) they are *$\mathcal{R}$-incomparable*.

The proof of the following Proposition (4.7) is omitted as the proof is contained in the proof of Proposition 4.8.

Proposition 4.7. *The relation $<_o$ is a strict partial order on $\mathbb{R}^+ \cup \{0\}$.*

Proposition 4.8. *The relation $<_c$ is a strict partial order on $\mathbb{R}^+ \cup \{0\}$.*

Proof. Irreflexivity follows directly from the definition and the irriflexivity of the relation $<$ on $\mathbb{R}^+ \cup \{0\}$. To show transitivity assume $[x_0, y_0] <_c [x_1, y_1]$ and $[x_1, y_1] <_c [x_2, y_2]$ for some $x_0, x_1, x_2, y_1, y_2, y_3 \in \mathbb{R}^+ \cup \{0\}$. If $x_0 < x_1$ and $y_0 \leq y_1$, or $x_0 \leq x_1$ and $y_0 < y_1$ then the result yields (transitivity holds). The remaining case for $[x_0, y_0] <_c [x_1, y_1]$ has $0 < x_0 - x_1 < y_1 - y_0$ therefore $y_0 < y_1 \leq y_2$. If $x_2 \geq x_0$ the result yields. Assume $x_2 < x_0$. The result yields by addressing the final cases $x_1 < x_2$ and $y_1 \leq y_2$, $x_1 \leq x_2$ and

$y_1 < y_2$, and $0 < x_1 - x_2 < y_2 - y_1$ under the assumptions: $x_2 < x_0$ and $0 < x_0 - x_1 < y_1 - y_0$.

Antisymmetry is implied by irreflexivity and transitivity. Hence $<_c$ is a strict partial order. □

The following result allows us to prove some classifying properties about dissimilarity maps. Proposition 4.9 shows the connection between the interval relations and the identity $x + y < w + z$ for $x, y, w, z \in \mathbb{R}^+ \cup \{0\}$. In the next chapter, it contributes to an indirect way of showing the four-point condition (Definition 3.18) holds for certain dissimilarity maps.

Proposition 4.9. *Let $x, y, w, z \in \mathbb{R}^+ \cup \{0\}$. If $[x, y] <_o [w, z]$ then $[x, y] <_c [w, z]$. If $[x, y] <_c [w, z]$ then $x + y < w + z$. More briefly $[x, y] <_o [w, z]$ implies $[x, y] <_c [w, z]$ which implies $x + y < w + z$.*

Proof. Consider $x, y, w, z \in \mathbb{R}^+ \cup \{0\}$. If $[x, y] <_o [w, z]$ then either $x < w$ and $y \leq z$, or $x \leq w$ and $y < z$ in either case $[x, y] <_c [w, z]$ holds.

If $[x, y] <_c [w, z]$ then $x < w$ and $y \leq z$, $x \leq w$ and $y < z$, or $0 < x - w < z - y$. In the first two cases the result is found by combining the inequalities. In the last case

$$\begin{aligned} x - w < z - y &\Rightarrow x + y < z + w \\ &\Rightarrow x + y < w + z \end{aligned}$$

□

4.3.1 Order interval relation properties

The next three results show the connection between strictly-increasing functions and the order interval relation. The first shows that $<_o$ is preserved by every strictly-increasing function (applied to each member of both intervals). Furthermore, it shows if $[x, y] <_o [w, z]$ doesn't hold, then there is no strictly-increasing function f such that $[f(x), f(y)] <_o [f(w), f(z)]$. This is to be expected as strictly-increasing functions preserve order and $<_o$ is based on solely on the ordering of the interval elements. The other two results (Lemma 4.11 and Proposition 4.12) deal with existence criteria for

a strictly-increasing function $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ with $f(0) = 0$ such that $f(x) + f(y) = f(w) + f(z)$.

Proposition 4.10. *Let $x, y, w, z \in \mathbb{R}^+ \cup \{0\}$ and $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ be a strictly-increasing function. The relation $[f(x), f(y)] <_o [f(w), f(z)]$ holds if and only if $[x, y] <_o [w, z]$.*

Proof. Consider $x, y, w, z \in \mathbb{R}^+ \cup \{0\}$. Assume $x < w$ and $y \leq z$. As f is strictly-increasing (and therefore preserves order) $f(x) < f(w)$ and $f(y) \leq f(z)$. Similarly, $x \leq w$ and $y < z$ imply $f(x) \leq f(w)$ and $f(y) < f(z)$. Hence $[x, y] <_o [w, z] \Rightarrow [f(x), f(y)] <_o [f(w), f(z)]$.

For the remaining implication ($[f(x), f(y)] <_o [f(w), f(z)] \Rightarrow [x, y] <_o [w, z]$), without loss of generality assume $f(x) < f(w)$ and $f(y) \leq f(z)$. Both $x \geq w$ and $y > z$ lead to contradictions of f being strictly-increasing. As both contradictions are essentially the same, detail is only provided for the contradiction of $x \geq w$. Assume $x \geq w$. If $x = w$ then $f(x) = f(w)$ contradicting $f(x) < f(w)$. Alternatively if $x > w$ then $f(x) > f(w)$ by the definition of strict increase (Definition 3.1). This again contradicts $f(x) < f(w)$. Hence $x < w$ and (by the similar but omitted argument) $y \leq z$. Hence

$$[f(x), f(y)] <_o [f(w), f(z)] \Rightarrow [x, y] <_o [w, z].$$

□

Lemma 4.11. *Let $x, y, w, z \in \mathbb{R}^+ \cup \{0\}$. If $[x, y]$ and $[w, z]$ are $<_o$-incomparable, then either*

1. *$x < w$ and $z < y$,*
2. *$w < x$ and $y < z$, or*
3. *$w = x$ and $y = z$.*

Proof. If $[x, y]$ and $[w, z]$ are $<_o$-incomparable, then by definition none of the following statements hold:

1. $x < w$ and $y \leq z$,
2. $x \leq w$ and $y < z$,

3. $w < x$ and $z \leq y$,

4. $w \leq x$ and $z < y$,

Consider the three possible cases $x < w$, $x = w$ and $x > w$.

If $x < w$, then by (the falsehood of) Statement 1 $z < y$. Hence $x < w$ and $z < y$.

If $x = w$, then $y = z$ by the falsehood of Statement 2 and Statement 4.

If $x > w$, then by Statement 3 $y < z$ and the result yields. □

Proposition 4.12. *Let $x, y, w, z \in \mathbb{R}^+ \cup \{0\}$. If $[x, y]$ and $[w, z]$ are $<_o$-incomparable, then there exists a strictly-increasing function $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ with $f(0) = 0$ such that $f(x) + f(y) = f(w) + f(z)$.*

Proof. By Lemma 4.11 without loss of generality only the case where $x < w$ and $z < y$ both hold and the case where $w = x$ and $y = z$ both hold need be considered. In the first case consider the $strictly - increasing$ function $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ linearly interpolated through the points

$\mathbb{R}^+ \cup \{0\}$	0	x	w	z	y
$f(\mathbb{R}^+ \cup \{0\})$	0	1	2	3	4

Note that it is assumed f is linearly interpolated through the defined points. The result yields as

$$f(x) + f(y) = 5 = f(w) + f(z).$$

In the second case the result yeilds for any strictly-increasing function $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ with $f(0) = 0$ as $f(w) = f(x)$ and $f(y) = f(z)$. □

The next example is a simple way in which the tools developed so far can be used. In the next chapter, the same tools will be used (in a very similar way) to prove that the four-point condition holds for certain dissimilarity maps.

Example 4.13. Consider the intervals $[1, 4]$, $[3, 5]$ and $[2, 7]$ depicted in Figure 4.4. By Proposition 4.12, as $[3, 5]$ and $[2, 7]$ are $<_o$-incomparable, there exists a strictly-increasing function $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ with $f(0) = 0$ such that $f(3) + f(5) = f(2) + f(7)$. Furthermore, as $[1, 4] <_o [3, 5]$, Proposition 4.10 shows that $[f(1), f(4)] <_o [f(3), f(5)]$. Hence, by Proposition 4.9, $f(1) + f(4) < f(3) + f(5)$.

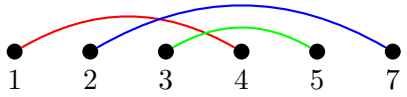


Figure 4.4: Three intervals: $[1, 4]$ (in red), $[3, 5]$ (in green) and $[2, 7]$ (in blue). The intervals $[3, 5]$ and $[2, 7]$ are $<_o$-incomparable. In contrast, the interval relation $[1, 4] <_o [3, 5]$ holds.

4.3.2 Convex interval relation properties

Proposition 4.14 and Corollary 4.15 are analogous to Proposition 4.10 using the convex interval relation $<_c$ instead of $<_o$. In particular, Proposition 4.14 shows that convex functions (applied to each element of an interval) preserve the convex interval relation.

Proposition 4.14. *Let $x, y, w, z \in \mathbb{R}^+ \cup \{0\}$. If the relation $[x, y] <_c [w, z]$ holds, then for all strictly-increasing convex functions $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$, the relation $[f(x), f(y)] <_c [f(w), f(z)]$ holds.*

Proof. If $x < w$ and $y \leq z$ or $x \leq w$ and $y < z$ then by Proposition 4.10 $[f(x), f(y)] <_o [f(w), f(z)]$ for all strictly-increasing functions $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ (including strictly-increasing convex functions). But by Proposition 4.9

$$[f(x), f(y)] <_o [f(w), f(z)] \Rightarrow [f(x), f(y)] <_c [f(w), f(z)]$$

so to complete the forward implication it suffices to show that for an arbitrary strictly-increasing convex function $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ if $0 < x - w < z - y$ then $0 < f(x) - f(w) < f(z) - f(y)$. As f is convex,

$$\frac{f(t) - f(a)}{t - a} \leq \frac{f(b) - f(t)}{b - t}$$

for all $t \in (a, b)$. In particular

$$\frac{f(x) - f(w)}{x - w} \leq \frac{f(y) - f(x)}{y - x}$$

and

$$\frac{f(y) - f(x)}{y - x} \leq \frac{f(z) - f(y)}{z - y}.$$

Combining the inequalities through the transitivity of $\leq$ yields

$$\frac{f(x) - f(w)}{x - w} \leq \frac{f(z) - f(y)}{z - y}.$$

But $0 < x - w < z - y$ so

$$\begin{aligned} f(x) - f(w) &\leq \frac{(f(z) - f(y))(x - w)}{z - y} \\ &< \frac{(f(z) - f(y))(x - w)}{x - w} \\ &= f(z) - f(y), \end{aligned}$$

hence $[f(x), f(y)] <_c [f(w), f(z)]$. □

Corollary 4.15. *Let $x, y, w, z \in \mathbb{R}^+ \cup \{0\}$. The relation $[x, y] <_c [w, z]$ holds if and only if, for all strictly-increasing convex functions $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$, the relation $[f(x), f(y)] <_c [f(w), f(z)]$ holds.*

Proof. The forward implication is given by Proposition 4.14.

For the backward implication assume for all strictly-increasing convex functions $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ the relation $[f(x), f(y)] <_c [f(w), f(z)]$ holds. Consider the strictly-increasing convex function $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ such that $f(t) = t$ for all $t \in \mathbb{R}^+ \cup \{0\}$. Then $[f(x), f(y)] = [x, y]$ and $[f(w), f(z)] = [w, z]$ and hence $[x, y] <_c [w, z]$. □

The following example is a typical use of Proposition 4.14.

Example 4.16. Consider the intervals $[1, 5]$ and $[2, 3]$. Let f be any strictly-increasing convex function. By Proposition 4.14, as $[2, 3] <_c [1, 5]$, the relation $[f(2), f(3)] <_c [f(1), f(5)]$ holds. Furthermore by Proposition 4.9, this implies $f(2) + f(3) < f(1) + f(5)$.

4.4 Clue functions

Clue functions are a special type of function that indicate whether other strictly-increasing convex functions, with certain properties, exist. Namely, given $x, y, w, z \in \mathbb{R}^+ \cup \{0\}$, if the clue function on x, y, w, z is a strictly-increasing convex function, then there exists a class of strictly-increasing convex functions with $f(0) = 0$ and $f(x) + f(y) = f(w) + f(z)$. However, if

the clue function on x, y, w, z is not strictly-increasing convex, then no such class of strictly-increasing convex functions exist. This fact is exploited in the next chapter.

Definition 4.17. Let $x, y, w, z \in \mathbb{R}^+ \cup \{0\}$ and without loss of generality assume $x \leq w \leq z \leq y$. The clue function $\zeta : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ on the set $\{x, y, w, z\}$ is the linear interpolation through the points

$\mathbb{R}^+ \cup \{0\}$	0	x	w	z	y
$\zeta(\mathbb{R}^+ \cup \{0\})$	0	x	w	z	$w + z - x$

.

It is possible for the clue function not to be well-defined in the case where the two largest elements are equal but the two smallest are not. For example clue function $\zeta : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ on the set $\{1, 2, 3, 3\}$ is not a function, as $\zeta(3) = 3$ and simultaneously $\zeta(3) = 4$. This warrants care be taken to ensure the clue function is well-defined before use. A particular case of interest is if some pairwise partition of $\{x, y, w, z\}$ is $<_c$-incomparable. In this case, the clue function on $\{x, y, w, z\}$ with $x, y, w, z \in \mathbb{R}^+ \cup \{0\}$ is well-defined as a consequence of Lemma 4.19.

Example 4.18. The clue function $\zeta : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ on the set $\{6, 4, 1, 5\}$ is well-defined. It is the linear interpolation through the points

$\mathbb{R}^+ \cup \{0\}$	0	1	4	5	6
$\zeta(\mathbb{R}^+ \cup \{0\})$	0	1	4	5	8

.

The next two lemmas are in preparation for Proposition 4.21, which relates strictly-increasing convex clue functions to the convex interval relation. It is the convex analogue to the order interval relation result Proposition 4.12.

Lemma 4.19. *Let $x, y, w, z \in \mathbb{R}^+ \cup \{0\}$. If $[x, y]$ and $[w, z]$ are incomparable by $<_c$, then either*

1. $0 < y - z \leq w - x$,
2. $0 < z - y \leq x - w$, *or*
3. $0 = y - z = w - x$.

Proof. If $[x, y]$ and $[w, z]$ are incomparable by $<_c$ then by definition none of the following statements hold:

1. $x < w$ and $y \leq z$,
2. $x \leq w$ and $y < z$,
3. $0 < x - w < z - y$,
4. $w < x$ and $z \leq y$,
5. $w \leq x$ and $z < y$,
6. $0 < w - x < y - z$.

Consider the three possible cases $x < w$, $x = w$ and $x > w$.

If $x < w$, then by (the falsehood of) Statement 1 $z < y$. Hence as $0 < w - x$ and $0 < y - z$, the inequality $0 < y - z \leq w - x$ holds by the falsehood of Statement 6.

If $x = w$, then $y = z$ by the falsehood of Statement 2 and Statement 5. In this case $0 = y - z = w - x$.

If $x > w$ then by Statement 4 $y < z$. Hence as $0 < x - w$ and $0 < z - y$, the inequality $0 < z - y \leq x - w$ holds by the falsehood of Statement 3. In each case the result holds. □

Lemma 4.20. *Let $x, y, w, z \in \mathbb{R}^+ \cup \{0\}$ and the relations $x \leq y$ and $w \leq z$ hold. Let $\zeta : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ be the clue function on $\{x, y, w, z\}$. If $0 < y - z \leq w - x$, then ζ is a strictly-increasing convex function and $\zeta(x) + \zeta(y) = \zeta(w) + \zeta(z)$.*

Proof. The clue function ζ with the above conditions is a strictly-increasing convex function as its linear components have successively increasing gradients (this is shown in depth in Example 3.6). Note, ζ is well defined as the premise stipulates $0 < y - z$ so $z < y$.

For the proposition to yield, the equality $\zeta(x) + \zeta(y) = \zeta(w) + \zeta(z)$ must be established. But

$$\zeta(x) + \zeta(y) = x + z + w - x = z + w = \zeta(w) + \zeta(z).$$

□

Proposition 4.21. *Let $x, y, w, z \in \mathbb{R}^+$. If $[x, y]$ and $[w, z]$ are incomparable by $<_c$, then the clue function $\zeta : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ on $\{x, y, w, z\}$ is a strictly-increasing convex function and $\zeta(x) + \zeta(y) = \zeta(w) + \zeta(z)$.*

Proof. By Lemma 4.19 either $0 < y - z \leq w - x$, $0 < z - y \leq x - w$, or $0 = y - z = w - x$. In the first two cases the result holds due to Lemma 4.20. In the last case ζ is well defined as $0 < x$ and $w = x$. Furthermore, $y = z$ and $w = x$ implies $\zeta(y) = \zeta(z)$ and $\zeta(w) = \zeta(x)$. The clue function is a strictly-increasing convex function as

$$\zeta(y) = w + z - x = x - x + y = y,$$

and therefore ζ is just the identity function. In each case the result yeilds. □

Example 4.22. Consider the clue function $\zeta : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ on the set $\{4, 3, 5, 1\}$. It is the linear interpolation through the points

$\mathbb{R}^+ \cup \{0\}$	0	1	3	4	5
$\zeta(\mathbb{R}^+ \cup \{0\})$	0	1	3	4	6

;

hence, it is well-defined. As the intervals $[1, 5]$ and $[3, 4]$ are $<_c$-incomparable, by Proposition 4.21, the function ζ is a strictly-increasing convex function.

The next result is the final for this chapter. It formalises the way in which clue functions act as a boundary for the class of strictly-increasing convex functions with $f(0) = 0$ and $f(x) + f(y) = f(w) + f(z)$. This class of functions is important for the next chapter as they are strictly-increasing convex transforms which lead to the four-point condition holding. The mean value theorem is used implicitly in the proof of Proposition 4.23.

Proposition 4.23. *Let $x, y, w, z \in \mathbb{R}^+$ such that $[x, y]$ and $[w, z]$ are incomparable by $<_c$ and $y \geq z$. Let $\zeta : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ be the clue function on $\{x, y, w, z\}$. Let $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ be a strictly-increasing convex function with $f(0) = 0$, $f(x) + f(y) = f(w) + f(z)$ and $f(y) = \zeta(y)$. If $\zeta(w) + \zeta(t_0) \leq \zeta(y)$ for some $t_0 \in \mathbb{R}^+ \cup \{0\}$, then $f(w) + f(t_0) \leq \zeta(y)$. Similarly, if $\zeta(z) + \zeta(t_1) \leq \zeta(y)$ for some $t_1 \in \mathbb{R}^+ \cup \{0\}$, then $f(z) + f(t_1) \leq \zeta(y)$.*

Proof. By Lemma 4.19 either $0 < y - z \leq w - x$, $0 < z - y \leq x - w$, or $0 = y - z = w - x$. In the case $0 = y - z = w - x$ the function ζ is a straight line through the points $(0, 0)$ and (y, y). Hence both the result holds trivially by the convexity of f (i.e. if $f(t) > \zeta(t)$ for any $0 < t < y$, then f would not be convex as $f(0) = 0$ and $f(y) = \zeta(y) = y$). Note ζ is well defined as

$$\zeta(y) = w + z - x = y + x - x = y \in \mathbb{R}^+.$$

Under the premise $y \geq z$, there is just one remaining case: $0 < y - z \leq w - x$. Assume $\zeta(w) + \zeta(t_0) \leq \zeta(y)$. By the convexity of ζ and Proposition 4.21 and as $\zeta(w)+\zeta(z) = \zeta(x)+\zeta(y)$ and $\zeta(w)+\zeta(t_0) \leq \zeta(y)$, the inequality $\zeta(t_0) \leq \zeta(z) - \zeta(x)$ holds. Therefore $t_0 \leq z - x$. Using the same rationale, under the assumption $\zeta(z) + \zeta(t_1) \leq \zeta(y)$, one can show $t_1 \leq w - x$.

As $f(0) = 0$ and f is a strictly-increasing convex function, $f(x) = s_1 x$, $f(w) = s_2 w$ and $f(z) = s_3 z$ for some non-decreasing list of scalars (s_1, s_2, s_3) such that $s_1, s_2, s_3 \in \mathbb{R}^+$. We begin by showing that each of the scalars s_i with $i \in \{1, 2\}$ have $s_i \leq 1$, as this implies $f(t) \leq \zeta(t)$ for all $t \in \mathbb{R}^+ \cup \{0\}$ with $0 \leq t \leq w$. To show $s_1 \leq 1$, note $f(y) = \zeta(y)$ and $f(x) + f(y) = f(w) + f(z)$ imply $s_1 x + w + z - x = s_2 w + s_3 z$. But

$$s_2 w + s_3 z \geq s_1 w + s_1 z = w + z + (s_1 - 1)(w + z)$$

and

$$s_1 x + w + z - x = w + z + (s_1 - 1)(x).$$

Hence $s_1 > 1$ implies

$$w + z + (s_1 - 1)(w + z) > w + z + (s_1 - 1)(x)$$

and leads to a contradiction of

$$s_1 x + w + z - x = s_2 w + s_3 z.$$

Alternatively assume $s_2 > 1$, then $s_3 > 1$. But this contradicts

$$s_1 x + w + z - x = s_2 w + s_3 z,$$

as $s_1 \leq 1$. Hence $s_2 \leq 1$.

We proceed by considering the structure of $f(t)$ for $t \in \mathbb{R}^+ \cup \{0\}$ with $t > w$. Let $\alpha, \beta \in \mathbb{R}^+ \cup \{0\}$ and $\gamma \in \mathbb{R}$ such that $f(x) = x - \alpha$, $f(w) = w - \beta$ and $f(z) = z - \gamma$. Note, α and β are non-negative by the previous paragraph. As

$$w - \beta + z - \gamma = f(w) + f(z) = f(x) + f(y) = x - \alpha + w + z - x = w + z - \alpha,$$

the equality $\gamma = \alpha - \beta$ is established. Hence $f(z) = z + \beta - \alpha$.

In this paragraph, we show the result for the $\zeta(w) + \zeta(t_0) \leq \zeta(y)$ case. Assume $\zeta(w) + \zeta(t_0) \leq \zeta(y)$. By the convexity of f and $t_0 \leq z$, the inequality

$$f(w) + f(t_0) = \zeta(w) - \beta + f(t_0) \leq \zeta(w) - \beta + \zeta(t_0) + \beta \leq \zeta(y)$$

holds. Note $f(t_0) \leq \zeta(t_0) + \beta$, else the gradient of f from t_0 to z would be less than that between w and t_0.

In this final paragraph, we assume $\zeta(z) + \zeta(t_1) \leq \zeta(y)$ and prove the remaining case. If $\beta \leq \alpha$, the result holds trivially, as $f(z) = z + \beta - \alpha \leq \zeta(z)$ and $f(t_1) < \zeta(t_1)$ (note $t_1 \leq w$ and $f(w) \leq \zeta(w)$ have both been established). Alternatively, if $\alpha < \beta$, by the convexity of f and $t_1 \leq w - x$, the inequality

$$f(z) + f(t_1) \leq \zeta z + \beta - \alpha + \zeta(t_1) + \alpha - \beta = \zeta(z) + \zeta(t_1) \leq \zeta(y)$$

holds. □

Chapter 5

Quartet classification and ultrametrics

In this final chapter, we classify four element dissimilarity maps by the sets of trees they are consistent with, under both the ordinal and convex assumptions. Four element dissimilarity maps are important as they correspond to phylogenetic X-trees with four leaves, which can be used to reconstruct larger trees. Indeed, as mentioned in Section 2.4, there are many reconstruction methods that exploit this fact (Strimmer and Von Haeseler, 1996; Roshan et al., 2004; Snir et al., 2008). One of these methods, the ordinal quartet method, has the ordinal assumption inbuilt and boasts an increased accuracy compared to similar methods based on the assumption of additivity (Kearney, 1998). The classification of dissimilarity maps presented in this chapter allows for the development of similar techniques based on the convex assumption.

In Section 5.1, the link between interval relations and dissimilarity maps is formalised. Classification of dissimilarity maps is presented under the ordinal assumption in Section 5.2, and under the convex assumption in Section 5.3. The final example from Section 3 is revisited and proven in Section 5.4. Finally, in Section 5.5, an independent proof of a Proposition on ultrametrics and the ordinal assumption is given.

5.1 Interval relations on dissimilarity maps

Proposition 5.1 narrows the set of possible phylogenetic X-trees that a dissimilarity map δ (on X) can fit, based on a certain interval relation holding. The set of possible fitted trees is narrowed to just one phylogenetic X-tree $\mathcal{T}$. Note, this does not guarantee that δ will fit $\mathcal{T}$.

Proposition 5.1. *Let $X = \{a, b, c, d\}$ and $[\delta(a, b), \delta(c, d)] <_o [\delta(a, c), \delta(b, d)]$ where δ is a dissimilarity map on X. Then, $\mathbf{T}_\delta^{o.e.} \subseteq \{ab|cd\}$. Similarly if $[\delta(a, b), \delta(c, d)] <_c [\delta(a, c), \delta(b, d)]$, then $\mathbf{T}_\delta^{c.r.} \subseteq \{ab|cd\}$.*

Proof. By applying both Proposition 4.14 and Proposition 4.9 to $[\delta(a, b), \delta(c, d)] <_c [\delta(a, c), \delta(b, d)]$ the following statement is proven: for all strictly-increasing convex functions f the inequality

$$f(\delta(a, b)) + f(\delta(c, d)) < f(\delta(a, c)) + f(\delta(b, d))$$

holds. The equivalent statement for strictly-increasing functions can be deduced from $[\delta(a, b), \delta(c, d)] <_o [\delta(a, c), \delta(b, d)]$ using Proposition 4.10 and Proposition 4.9. For the remainder of the proof, the arguments for both relation types are equivalent so only the convex case is given.

Suppose δ is convex related to some $\mathcal{T}$-metric $d_{(\mathcal{T},w)}$ for some weighted phylogenetic X-tree $(\mathcal{T}, w)$, then by the previous statement and the definition of convex relation

$$d_{(\mathcal{T},w)}(a, b) + d_{(\mathcal{T},w)}(c, d) < d_{(\mathcal{T},w)}(a, c) + d_{(\mathcal{T},w)}(b, d).$$

Hence by the four-point classification of $\mathcal{T}$ (Theorem 3.21) the $\mathcal{T}$-metric $d_{(\mathcal{T},w)}$ has type $ab|cd$. Therefore if δ is convex related to a $\mathcal{T}$-metric, then $\mathcal{T}$ is isomorphic to $ab|cd$ (by Proposition 3.20). This yields the result as $\mathbf{T}_\delta^{c.r.} \subseteq \{ab|cd\}$.

□

An astute reader may notice the proof and declaration of Proposition 5.1 have an implied assumption that was not addressed. To use the notation $[\delta(a, b), \delta(c, d)]$ for $X = \{a, b, c, d\}$ and with δ being a dissimilarity map on X, one is assuming that $\delta(a, b) \leq \delta(c, d)$. Obviously this assumption

is justified by a simple relabelling of the elements of X. But what of the assumption $\delta(a,c) \leq \delta(b,d)$ after said relabelling? Suppose instead that $\delta(a,c) \geq \delta(b,d)$. The relabelling $a \leftrightarrow b$ and $c \leftrightarrow d$ has no effect on the first pairwise partition but now $\delta(a,c) \leq \delta(b,d)$. Hence both are justified without loss of generality. There is one more pairwise partition $[\delta(a,d), \delta(b,c)]$ on X. But the assumption $\delta(a,d) \leq \delta(b,c)$ cannot be justified without loss of generality as the labelling is fixed by the previous two pairwise partitions. This means of the three pairwise partitions of X, the first two fix the labelling and the third must be considered with two cases. To address this issue, the notation

$$[x,y]' = \begin{cases} [x,y], & \text{if } x \leq y \\ [y,x], & \text{if } y < x \end{cases}$$

is introduced (in this case one would use the substitutions $x = \delta(a,d)$ and $y = \delta(b,c)$).

Before moving on, it is important to give credit to Kearney (1998). The justification for Kearney's ordinal quartet method is essentially an informal form of Proposition 5.1 (in the $<_o$ case) and is where the idea of using interval relations on elements of a dissimilarity map in this way originates. Although Kearney does not extend the idea to the types of classifications found in the next few sections, it is important to acknowledge that those extensions would not be possible without his contribution of the original idea.

Example 5.2. Let $X = \{a, b, c, d\}$ and

$\delta =$

	a	b	c	d
a	0	1	2	3
b	1	0	4	6
c	2	4	0	5
d	3	6	5	0

be a strong dissimilarity map on X. An important precursor to classification is identifying which interval relations hold on the pairwise partitions of X. By pairwise partitions of X, we are taking some creative liberty and actually referring to the intervals consisting of elements from the dissimilarity map δ,

corresponding to the pairwise partitions of X. More explicitly, the pairwise partitions of X refers to the intervals $[\delta(a,b),\delta(c,d)]'$, $[\delta(a,c),\delta(b,d)]'$ and $[\delta(a,d),\delta(b,c)]'$. Figure 5.1 depicts each of the pairwise partitions of X in this case.

As $[\delta(a,b),\delta(c,d)] <_o [\delta(a,c),\delta(b,d)]$, Proposition 5.1 shows $\mathbf{T}_\delta^{o.e.} \subseteq \{ab|cd\}$. Similarly, as $[\delta(a,b),\delta(c,d)] <_c [\delta(a,c),\delta(b,d)]$ and $[\delta(a,d),\delta(b,c)] <_c [\delta(a,c),\delta(b,d)]$, Proposition 5.1 can be applied twice to show $\mathbf{T}_\delta^{c.r.} = \emptyset$.

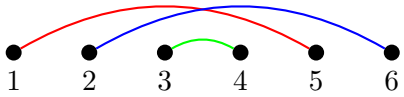


Figure 5.1: The interval $[1,5]$ (in red) corresponds to the interval $[\delta(a,b),\delta(c,d)]$. Similarly, the interval $[2,6]$ (in blue) corresponds to $[\delta(a,c),\delta(b,d)]$ and $[3,4]$ (in green) corresponds to $[\delta(a,d),\delta(b,c)]$.

5.2 Ordinal classification

In Definition 3.12 a strong dissimilarity map was introduced as a dissimilarity map $\delta : X \times X \to \mathbb{R}^+ \cup \{0\}$ with $\delta(a,b) = 0$ only if $a = b$ for all distinct $a, b \in X$. If this condition fails, then some distinct $a, b \in X$ have $\delta(a,b) = 0$. But any strictly-increasing function f between dissimilarity maps has $f(0) = 0$ and therefore $f(\delta(a,b)) = 0$. So $\mathbf{T}_\delta^{o.e.} = \emptyset$ as no phylogenetic X-tree weight function can have weight zero between leaves. Therefore a dissimilarity map δ needs to be a strong dissimilarity map, if $\mathbf{T}_\delta^{o.e.}$ is to be non-empty. With this in mind, Proposition 5.4 classifies $\mathbf{T}_\delta^{o.e.}$ for any strong dissimilarity map δ with at least one pairwise partition of X being $<_o$-comparable to another. This is in contrast to Proposition 5.1, as it gives the exact set $\mathbf{T}_\delta^{o.e.}$ (not just a super set).

Lemma 5.3. *Let $X = \{a,b,c,d\}$. Let δ_1 and δ_2 be strong dissimilarity maps on X such that $\delta_1 \overset{o.e.}{\sim} \delta_2$. If $\delta_2(a,b) + \delta_2(c,d) < \delta_2(a,c) + \delta_2(b,d)$ and $\delta_2(a,c) + \delta_2(b,d) = \delta_2(a,d) + \delta_2(b,c)$ hold, then $\{ab|cd\} \in \mathbf{T}_{\delta_1}^{o.e.}$. Similarly, if*

$$\delta_2(a,b) + \delta_2(c,d) = \delta_2(a,c) + \delta_2(b,d) = \delta_2(a,d) + \delta_2(b,c)$$

holds, then $\{abcd\} \in \mathbf{T}^{o.e.}_{\delta_1}$

Proof. Suppose $\delta_1 \overset{o.e.}{\sim} \delta_2$. Then there exists a strictly-increasing function $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ such that $\delta_2 = f(\delta_1)$. Let t_0 be a maximal value in the range of δ_2. Let $f_{t_0} : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ be the strictly-increasing function defined by $f_{t_0}(t) = f(t) + t_0$ for all $t \in \mathbb{R}^+$ and $f_{t_0}(0) = 0$. Let $\delta' : X \times X \to \mathbb{R}^+ \cup \{0\}$ be the dissimilarity map defined by $\delta' = f_{t_0}(\delta_1)$. Note δ' is a strong dissimilarity map as $f_{t_0}(0) = 0$ and f_{t_0} is strictly-increasing. If ,

$$\delta'(a,b) + \delta'(c,d) < \delta'(a,c) + \delta'(b,d)$$

and

$$\delta'(a,c)) + \delta'(b,d) = \delta'(a,d) + \delta'(b,c)$$

as neither identity is changed by the addition of t_0 to both sides. Similarly,

$$\delta'(a,c) + \delta'(b,d) = \delta'(a,c)) + \delta'(b,d) = \delta'(a,d) + \delta'(b,c)$$

To show the strict (with $<$ not just $\leq$) triangle inequality holds, consider any distinct $\alpha, \beta, \gamma \in X$. As δ' is a strong dissimilarity map, each of $\delta'(\alpha,\beta)$, $\delta'(\alpha,\gamma)$ and $\delta'(\gamma,\beta)$ are non-zero. But by the choice of t_0, $\delta'(\alpha,\beta) \leq 2t_0$ and $\delta'(\alpha,\gamma) + \delta'(\gamma,\beta) > 2t_0$ and so the triangle inequality holds.

Therefore the four-point condition holds and δ' has type $ab|cd$. Furthermore, $\delta_1 \overset{o.e.}{\sim} \delta'$ as f_{t_0} is strictly-increasing. Hence $\{ab|cd\} \in \mathbf{T}^{o.e.}_{\delta_1}$ by Proposition 3.20. □

Proposition 5.4. *Let* $X = \{a,b,c,d\}$ *and* $[\delta(a,b), \delta(c,d)] <_o [\delta(a,c), \delta(b,d)]$ *where* δ *is a strong dissimilarity map on* X. *If* $[\delta(a,c), \delta(b,d)]$ *and* $[\delta(a,d), \delta(b,c)]'$ *are* $<_o$*-comparable, then* $\mathbf{T}^{o.e.}_{\delta} = \emptyset$. *If* $[\delta(a,c), \delta(b,d)]$ *and* $[\delta(a,d), \delta(b,c)]'$ *are* $<_o$*-incomparable, then* $\mathbf{T}^{o.e.}_{\delta} = \{ab|cd\}$.

Proof. Let $X = \{a,b,c,d\}$ and

$$[\delta(a,b), \delta(c,d)] <_o [\delta(a,c), \delta(b,d)]$$

where δ is a strong dissimilarity map on X. Suppose $[\delta(a,c), \delta(b,d)]$ and $[\delta(a,d), \delta(b,c)]'$ are $<_o$-comparable. Without loss of generality assume

$$[\delta(a,c), \delta(b,d)] <_o [\delta(a,d), \delta(b,c)]',$$

then by Proposition 5.1 $\mathbf{T}_\delta^{o.e.} \subseteq \{ab|cd\}$ and $\mathbf{T}_\delta^{o.e.} \subseteq \{ac|bd\}$. Hence if $[\delta(a,c), \delta(b,d)]$ and $[\delta(a,d), \delta(b,c)]'$ are $<_o$-comparable, then $\mathbf{T}_\delta^{o.e.} = \emptyset$.

For the remaining claim, suppose $[\delta(a,c), \delta(b,d)]$ and $[\delta(a,d), \delta(b,c)]'$ are $<_o$-incomparable. By Proposition 4.12 there exists a strictly-increasing function $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ with $f(0) = 0$ and

$$f(\delta(a,c)) + f(\delta(b,d)) = f(\delta(a,d)) + f(\delta(b,c)).$$

Moreover, by Proposition 4.14 and Proposition 4.9

$$f(\delta(a,b)) + f(\delta(c,d)) < f(\delta(a,c)) + f(\delta(b,d)).$$

Let $\delta_f = f(\delta)$. Note δ_f is a strong dissimilarity map as f is strictly-increasing and $f(0) = 0$. Hence by Lemma 5.3 and Proposition 5.1 $\mathbf{T}_\delta^{o.e.} = \{ab|cd\}$ and the result yeilds. □

The advantages of Proposition 5.4 over Proposition 5.1 are made explicit in the following example.

Example 5.5. Let $X = \{a, b, c, d\}$ and

$\delta =$

	a	b	c	d
a	0	1	2	3
b	1	0	4	6
c	2	4	0	5
d	3	6	5	0

be the strong dissimilarity map from Example 5.2. Refer back to Figure 5.1 for a depiction of the interval $[\delta(a,b), \delta(c,d)]$, $[\delta(a,c), \delta(b,d)]$ and $[\delta(a,d), \delta(b,c)]$ (which we refer to as the pairwise partitions of X). In Example 5.2, as $[\delta(a,b), \delta(c,d)] <_o [\delta(a,c), \delta(b,d)]$, we used Proposition 5.1 to show $\mathbf{T}_\delta^{o.e.} \subseteq \{ab|cd\}$. But, as $[\delta(a,c), \delta(b,d)]$ and $[\delta(a,d), \delta(b,c)]$ are $<_o$-incomparable, Proposition 5.4 proves $\mathbf{T}_\delta^{o.e.} = \{ab|cd\}$.

To complete the classification of an arbitrary dissimilarity map δ on four elements under the ordinal assumption, the cases where no pairwise partitions of X are comparable by $<_o$ must be considered. These cases are difficult, as Proposition 5.1 cannot be used to narrow the set of phylogenetic X-trees that fit δ.

Proposition 5.6. *Let $X = \{a, b, c, d\}$ and let δ be a strong dissimilarity map on X such that $[\delta(a,b), \delta(c,d)]$, $[\delta(a,c), \delta(b,d)]$ and $[\delta(a,d), \delta(b,c)]'$ are all $<_o$-incomparable with each other. If $[\delta(a,b), \delta(c,d)]$, $[\delta(a,c), \delta(b,d)]$ and $[\delta(a,d), \delta(b,c)]'$ are distinct, then $\mathbf{T}_\delta^{o.e.} = \{ab|cd, ac|bd, ad|bc, abcd\}$. Alternatively, if $[\delta(a,c), \delta(b,d)]$ and $[\delta(a,d), \delta(b,c)]'$ are the same but $[\delta(a,b), \delta(c,d)]$ is distinct then $\mathbf{T}_\delta^{o.e.} = \{ab|cd, abcd\}$. Lastly, if $[\delta(a,b), \delta(c,d)]$, $[\delta(a,c), \delta(b,d)]$ and $[\delta(a,d), \delta(b,c)]'$ are all the same then $\mathbf{T}_\delta^{o.e.} = \{abcd\}$.*

Proof. Some preliminaries, common in the proof of each case, are introduced in this paragraph. Each case is shown individually in the proceeding paragraphs. As such, all assignments made in this paragraph are retained for the whole proof and all assignments made in successive paragraphs are only retained for that paragraph. By Proposition 4.12, as $[\delta(a,b), \delta(c,d)]$, $[\delta(a,c), \delta(b,d)]$ and $[\delta(a,d), \delta(b,c)]'$ are all $<_o$-incomparable to one another, there exist strictly-increasing functions $f, g, h : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ such that

$$f(\delta(a,c)) + f(\delta(b,d)) = f(\delta(a,d)) + f(\delta(b,c)),$$

$$g(\delta(a,b)) + g(\delta(c,d)) = g(\delta(a,d)) + g(\delta(b,c)),$$

$$h(\delta(a,b)) + h(\delta(c,d)) = h(\delta(a,c)) + h(\delta(b,d))$$

and $f(0) = g(0) = h(0) = 0$. When pairwise partitions of X are mentioned in this proof, we are referring to $[\delta(a,b), \delta(c,d)]$, $[\delta(a,c), \delta(b,d)]$ and $[\delta(a,d), \delta(b,c)]'$. Because of Proposition 2.23, we only need to consider the phylogenetic X-trees $ab|cd$, $ac|bd$, $ad|bc$ and $abcd$.

This paragraph deals with the case where each pairwise partition of X is distinct. Assume the pairwise partitions of X are distinct and let the multiset S be defined by

$$S = \{f(\delta(a,b)), f(\delta(c,d)), f(\delta(a,c)), f(\delta(b,d)), f(\delta(a,d)), f(\delta(b,c))\}.$$

Note, we call S a multiset, as Lemma 4.11 cannot rule out the case where one of the three intervals is trivial (e.g. $\delta(a,c) = \delta(b,d)$); however, Lemma 4.11 and the assumption that each interval is distinct do prove that every element of S, with the exception of those from at most one of the intervals, is distinct. Let $L = (x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7)$ be a increasing ordering

of $S \cup \{0, x_7\}$ with $x_0 = 0$ and $x_7 = 2x_6$. As the pairwise partitions of X are distinct and $<_o$-incomparable, by Lemma 4.11, the list L must be in one of the two forms shown in Figure 5.2. For each element $x \in S$, let $\underline{x} \in S \cup 0$ denote the next smallest term in L and $\overline{x} \in S \cup x_7$ denote the next biggest term in L. More formally, if $x = x_i \in S$, then $\underline{x} = x_{i-1}$ unless $x = x_{i-1}$, in which case $\underline{x} = x_{i-2}$ (note this can only occur if $x = x_4$, so x_{i-2} is well defined). Similarly, if $x = x_i$, then $\overline{x} = x_{i+1}$ unless $x = x_{i+1}$, in which case $\overline{x} = x_{i+2}$ (again this is well defined as this can only occur if $x = x_3$). Let ϵ be half the minimum positive distance between any two elements of L. Consider the function $\underline{f} : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$, such that $\underline{f}$ is a linear interpolation through a non-decreasing ordering (in terms of the first column) of the points

$\mathbb{R}^+ \cup \{0\}$	$\underline{f}(\mathbb{R}^+ \cup \{0\})$
0	0
$\delta(a,b)$	$\underline{f(\delta(a,b))} + \epsilon$
$\delta(c,d)$	$\underline{f(\delta(c,d))} + \epsilon$
$\delta(a,c)$	$f(\delta(a,c))$
$\delta(b,d)$	$f(\delta(b,d))$
$\delta(a,d)$	$f(\delta(a,d))$
$\delta(b,c)$	$f(\delta(b,c))$

.

By the choice of ϵ and Lemma 4.11

$$\underline{f}(\delta(a,b)) + \underline{f}(\delta(c,d)) < \underline{f}(\delta(a,c)) + \underline{f}(\delta(b,d)) = \underline{f}(\delta(a,d)) + \underline{f}(\delta(b,c)).$$

But the function $\underline{f}$ is order-preserving (also by choice of ϵ), hence $\delta \overset{o.e.}{\sim} \underline{f}(\delta)$ and therefore $ab|cd \in \mathbf{T}_\delta^{o.e.}$ by Lemma 5.3. The same argument can be made with g and h to show $ac|bd, ad|bc \in \mathbf{T}_\delta^{o.e.}$. To finish this case, it remains to show that $abcd \in \mathbf{T}_\delta^{o.e.}$. The way in which we do this is reminiscent of Example 3.9, where a desired function is found by mediating two extremes. To this end, consider the function $\overline{f} : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$, such that $\overline{f}$ is a linear interpolation through a non-decreasing ordering (in terms of the first

column) of the points

$\mathbb{R}^+ \cup \{0\}$	$\overline{f}(\mathbb{R}^+ \cup \{0\})$
0	0
$\delta(a,b)$	$\overline{f(\delta(a,b))} - \epsilon$
$\delta(c,d)$	$\overline{f(\delta(c,d))} - \epsilon$
$\delta(a,c)$	$f(\delta(a,c))$
$\delta(b,d)$	$f(\delta(b,d))$
$\delta(a,d)$	$f(\delta(a,d))$
$\delta(b,c)$	$f(\delta(b,c))$

.

By the choice of ϵ and Lemma 4.11

$$\overline{f}(\delta(a,b)) + \overline{f}(\delta(c,d)) > \overline{f}(\delta(a,c)) + \overline{f}(\delta(b,d)) = \overline{f}(\delta(a,d)) + \overline{f}(\delta(b,c)).$$

The function $\overline{f}$ is also order-preserving (by the choice of ϵ). Let the function $\mu : \mathbb{R} \to \mathbb{R}$ be defined by

$$\mu(t) = (1-t)\underline{f}(\delta(a,b)) + t\overline{f}(\delta(a,b)) + (1-t)\underline{f}(\delta(c,d)) + t\overline{f}(\delta(c,d)).$$

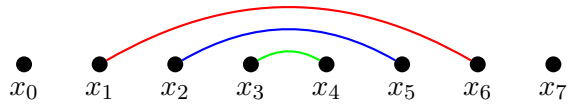
As $\mu(0) < f(\delta(a,c)) + f(\delta(b,d))$ and $\mu(1) > f(\delta(a,c)) + f(\delta(b,d))$, by the intermediate value theorem and the continuity of μ (with respect to t) there exists a $t_0 \in \mathbb{R}^+ \cup \{0\}$ such that $\mu(t_0) = f(\delta(a,c)) + f(\delta(b,d))$. Hence, the function $f_{t_0} : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ defined by $f_{t_0}(s) = (1-t_0)\underline{f}(s) + t_0\overline{f}(s)$ for all $s \in \mathbb{R}^+ \cup \{0\}$ has

$$f_{t_0}(\delta(a,b)) + f_{t_0}(\delta(c,d)) = f_{t_0}(\delta(a,c)) + f_{t_0}(\delta(b,d)) = f_{t_0}(\delta(a,d)) + f_{t_0}(\delta(b,c)).$$

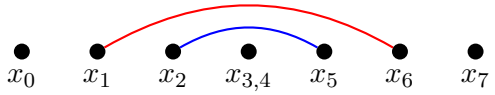
Furthermore, the function f_{t_0} is strictly-increasing (by Proposition 3.7 and Proposition 3.8) and $f_{t_0}(0) = 0$. Hence $\delta \overset{o.e.}{\sim} \underline{f}(\delta)$ and therefore $abcd \in \mathbf{T}_\delta^{o.e.}$, by Lemma 5.3. Therefore, we have shown $\mathbf{T}_\delta^{o.e.} = \{ab|cd, ac|bd, ad|bc, abcd\}$ (there are no other phylogenetic 4-trees to consider).

In this paragraph, assume $[\delta(a,c), \delta(b,d)]$ and $[\delta(a,d), \delta(b,c)]'$ are the same but $[\delta(a,b), \delta(c,d)]$ is distinct. Then, for any strictly-increasing function $\hat{f} : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$, the equation

$$\hat{f}(\delta(a,c)) + \hat{f}(\delta(b,d)) = \hat{f}(\delta(a,d)) + \hat{f}(\delta(b,c))$$



(a) Case with $x_3 < x_4$.



(b) Case with $x_3 = x_4$.

Figure 5.2: Both possible cases for the ordering of the points in the list $L = \{x_0, x_1, x_2 \dots x_7\}$ and the intervals $[x_1, x_6]$ (in red), $[x_2, x_5]$ (in blue) and $[x_3, x_4]$ (in green), which correspond to the pairwise partition of X. Note that $x_0 = 0$ and $x_7 = 2x_6$

holds and hence $\hat{f}(\delta)$ cannot have type $ac|bd$ or $ad|bc$ (assuming it passes the four-point condition and is a strong dissimilarity map). Hence, $ac|bd$ and $ad|bc$ are not contained in $\mathbf{T}_\delta^{o.e.}$. However, g is a strictly-increasing function with $g(0) = 0$ and

$$g(\delta(a,b)) + g(\delta(c,d)) = g(\delta(a,c)) + g(\delta(b,d)) = g(\delta(a,d)) + g(\delta(b,c)).$$

Hence, as $\delta \overset{o.e.}{\sim} f(\delta)$ and by Lemma 5.3, the inclusion $abcd \in \mathbf{T}_\delta^{o.e.}$ holds. It remains to show $ab|cd \in \mathbf{T}_\delta^{o.e.}$. This is done using a simplified version of an argument from the previous paragraph. As such, the details are explained more briefly. Let the multiset S be defined by $S = \{f(\delta(a,b)), f(\delta(c,d)), f(\delta(a,d)), f(\delta(b,c))\}$. Let $L = (x_0, x_1, x_2, x_3, x_4)$ be a non-decreasing ordering of $S \cup \{0\}$ with $x_0 = 0$. For each element $x \in S$, let $\underline{x} \in S \cup 0$ denote the next smallest term in L. Let ϵ be half the minimum positive distance between any two elements of L. Consider the function $\underline{f} : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$, such that $\underline{f}$ is a linear interpolation through a

non-decreasing ordering (in terms of the first column) of the points

$\mathbb{R}^+ \cup \{0\}$	$\underline{f}(\mathbb{R}^+ \cup \{0\})$
0	0
$\delta(a,b)$	$\underline{f(\delta(a,b))} + \epsilon$
$\delta(c,d)$	$\underline{f(\delta(c,d))} + \epsilon$
$\delta(a,d)$	$f(\delta(a,d))$
$\delta(b,c)$	$f(\delta(b,c))$

.

By the choice of ϵ and Lemma 4.11

$$\underline{f}(\delta(a,b)) + \underline{f}(\delta(c,d)) < \underline{f}(\delta(a,d)) + \underline{f}(\delta(b,c)).$$

The function $\underline{f}$ is order-preserving and $[\delta(a,c), \delta(b,d)] = [\delta(a,d), \delta(b,c)]'$; hence, $\delta \overset{o.e.}{\sim} \underline{f}(\delta)$ and therefore $ab|cd \in \mathbf{T}_{\delta}^{o.e.}$ by Lemma 5.3.

In this final paragraph, all the pairwise partitions of X are assumed to be the same. Hence, f is a strictly-increasing function with $f(0) = 0$ and

$$f(\delta(a,b)) + f(\delta(c,d)) = f(\delta(a,c)) + f(\delta(b,d)) = f(\delta(a,d)) + f(\delta(b,c)).$$

As $\delta \overset{o.e.}{\sim} f(\delta)$ and by Lemma 5.3, the inclusion $abcd \in \mathbf{T}_{\delta}^{o.e.}$ holds. Note, $f(\delta)$ is a strong dissimilarity map as $f(0) = 0$ and δ is a strong dissimilarity map. The phylogenetic X-trees $ab|cd$, $ac|bd$ and $ad|bc$ are not contained in $\mathbf{T}_{\delta}^{o.e.}$, as for any strictly-increasing function $\hat{f} : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ the equation

$$\hat{f}(\delta(a,b)) + \hat{f}(\delta(c,d)) = \hat{f}(\delta(a,c)) + \hat{f}(\delta(b,d)) = \hat{f}(\delta(a,d)) + \hat{f}(\delta(b,c))$$

holds and hence $\hat{f}(\delta)$ can only have type $abcd$ (if it even passes the four-point condition and is a strong dissimilarity map). □

Example 5.7. Let $X = \{a, b, c, d\}$ and

$\delta =$

	a	b	c	d
a	0	3	1	1
b	3	0	8	8
c	1	8	0	5
d	1	8	5	0

be a strong dissimilarity map on X. Figure 5.3 depicts each of the pairwise partitions of X. As all the pairwise partitions of X are $<_o$-incomparable, Proposition 5.6 can be applied. Combined with the fact that $[\delta(a,c),\delta(b,d)] = [\delta(a,d),\delta(b,c)]$, this proves that $\mathbf{T}_\delta^{o.e.} = \{ab|cd, abcd\}$.

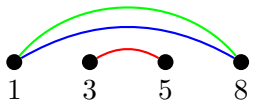


Figure 5.3: The interval $[3,5]$ (in red) corresponds to the interval $[\delta(a,b),\delta(c,d)]$. Similarly, the interval $[1,8]$ (in both blue and green) corresponds to $[\delta(a,c),\delta(b,d)]$ and $[\delta(a,d),\delta(b,c)]$.

5.3 Convex classification

Just as strong dissimilarity maps were a starting point for classification under the interval relation, strict metrics are required for classification under $<_c$. Proposition 5.9 formalises this notion. Proposition 5.9 proves, for $\mathbf{T}_\delta^{c.r.}$ to be non-empty for some dissimilarity map δ, that dissimilarity map must be a metric. One can further show, using a similar argument, that δ must be a strict metric by Proposition 3.17, as $\mathbf{T}_\delta^{c.r.}$ is a set of phylogenetic X-trees and not other types of X-trees.

Lemma 5.8. *Let $x, y, z \in \mathbb{R}^+ \cup \{0\}$. If $x + y < z$ holds, then for all strictly-increasing convex functions $f : \mathbb{R} \to \mathbb{R}$, the relation $f(x)+f(y) < f(0)+f(z)$ holds.*

Proof. Assume without loss of generality $x \leq y$. The relation $[x,y] <_c [0,z]$ holds as $x - 0 < z - y$. Hence, by Proposition 4.14 for all strictly-increasing convex functions $f : \mathbb{R} \to \mathbb{R}$, the relation $[f(x), f(y)] <_c [f(0), f(z)]$ holds. The result follows by Proposition 4.9. □

Proposition 5.9. *If δ is a dissimilarity map on X and the triangle inequality for δ does not hold, then $\mathbf{T}_\delta^{c.r.} = \emptyset$. Furthermore, δ fits no X-trees under the convex relation.*

Proof. Let δ be a dissimilarity map on X that fails the triangle inequality. Therefore there exist $a, b, c \in X$ such that $\delta(a,b) + \delta(a,c) < \delta(b,c)$. Let $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ be any strictly-increasing convex function with $f(0) = 0$. Let δ_f be the dissimilarity map obtained by setting $\delta_f(\alpha, \beta) = f(\delta(\alpha, \beta))$ for all $\alpha, \beta \in X$. Note $f(0) = 0$ is required, as otherwise δ_f is not a dissimilarity map. By Lemma 5.8 $f(\delta(a,b)) + f(\delta(a,c)) < f(\delta(b,c)) + f(0)$ which implies $\delta_f(a,b) + \delta_f(a,c) < \delta_f(b,c)$. So δ_f fails the triangle inequality and therefore cannot be a tree metric (Proposition 3.21). As the choice of f was arbitrary δ fits no X-tree under convex relation. □

The partial classification of an arbitrary strict metric δ on four elements is presented in two parts. The first part, Proposition 5.10, narrows the set of trees that δ fits to the empty set, in the case where certain pairwise partitions of X are $<_c$-comparable. The second part, Proposition 5.11, classifies all cases not treated in Proposition 5.10 with at least one pairwise partition of X being $<_c$-comparable. It does not provide a full classification as it does not consider the case when no pairwise partitions of X are comparable by $<_c$. It should be mentioned that despite this being a partial classification, it is still enough to develop a classification algorithm. Indeed, even the ordinal quartet method (Kearney, 1998) relied simply on the equivalent of Proposition 5.1 (a much simpler classification then the following). The ordinal quartet method classifies dissimilarity maps, that cannot be classified under Kearney's partial ordinal classification, by reverting to another classification method, just for those (preferably rare) inputs.

Proposition 5.10. *Let* $X = \{a, b, c, d\}$ *and* $[\delta(a,b), \delta(c,d)] <_c [\delta(a,c), \delta(b,d)]$ *where* δ *is a strict metric on* X. *If* $[\delta(a,c), \delta(b,d)]$ *and* $[\delta(a,d), \delta(b,c)]'$ *are* $<_c$*-comparable, then* $\mathbf{T}_\delta^{c.r.} = \emptyset$.

Proof. Without loss of generality assume

$$[\delta(a,c), \delta(b,d)] <_c [\delta(a,d), \delta(b,c)]'.$$

By Proposition 5.1 $\mathbf{T}_\delta^{c.r.} \subseteq \{ac|bd\}$. But by the premise

$$[\delta(a,b), \delta(c,d)] <_c [\delta(a,c), \delta(b,d)],$$

Proposition 5.1 also yeilds $\mathbf{T}_\delta^{c.r.\cdot} \subseteq \{ab|cd\}$. Hence $\mathbf{T}_\delta^{c.r.\cdot} = \emptyset$. Note there was no loss of generality as the argument would be the same under the assumption

$$[\delta(a,d),\delta(b,c)]' <_c [\delta(a,c),\delta(b,d)].$$

□

Proposition 5.11. *Let* $X = \{a,b,c,d\}$ *and* $[\delta(a,b),\delta(c,d)] <_c [\delta(a,c),\delta(b,d)]$ *where* δ *is a strict metric on* X*. Suppose* $[\delta(a,c),\delta(b,d)]$ *and* $[\delta(a,d),\delta(b,c)]'$ *are* $<_c$*-incomparable. Let* $\zeta : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ *be the clue function on* $\{\delta(a,c),\delta(a,d),\delta(b,c),\delta(b,d)\}$*. If* $\zeta(\delta)$ *is a strict metric, then* $\mathbf{T}_\delta^{c.r.\cdot} = \{ab|cd\}$*. If* $\zeta(\delta)$ *is not a strict metric, then* $\mathbf{T}_\delta^{c.r.\cdot} = \emptyset$*.*

Proof. Firstly, note the clue function ζ is well defined as $[\delta(a,c),\delta(b,d)]$ and $[\delta(a,d),\delta(b,c)]'$ are $<_c$-incomparable. Assume $\delta_\zeta = \zeta(\delta)$ is a strict metric. By Proposition 4.21, ζ is convex and

$$\delta_\zeta(a,c) + \delta_\zeta(b,d) = \delta_\zeta(a,d) + \delta_\zeta(b,c).$$

As

$$[\delta(a,b),\delta(c,d)] <_c [\delta(a,c),\delta(b,d)],$$

by Proposition 4.14 and Proposition 4.9

$$\delta_\zeta(a,b) + \delta_\zeta(c,d) < \delta_\zeta(a,c) + \delta_\zeta(b,d).$$

But δ_ζ is also a strict metric so the four-point condition holds and δ_ζ has type $ab|cd$. Therefore δ_ζ is an $ab|cd$-metric by Proposition 3.20. Hence by Proposition 5.1, $\mathbf{T}_\delta^{c.r.\cdot} = \{ab|cd\}$.

Now suppose δ_ζ is not a strict metric and let the list (x,w,z,y) be a non-decreasing ordering of $(\delta(a,c),\delta(a,d),\delta(b,c),\delta(b,d))$. Note that as $[\delta(a,c),\delta(b,d)]$ and $[\delta(a,d),\delta(b,c)]'$ are $<_c$-incomparable

$$\{x,y\},\{w,z\} \in \{\{\delta(a,d),\delta(b,c)\},\{\delta(a,c),\delta(b,d)\}\}.$$

As δ_ζ is not a strict metric there exist elements $\alpha,\beta,\gamma \in X$ such that

$$\zeta(\delta(\alpha,\beta)) + \zeta(\delta(\beta,\gamma)) \leq \zeta(\delta(\alpha,\gamma)).$$

By the definition of ζ and

$$[\delta(a,b),\delta(c,d)] <_c [\delta(a,c),\delta(b,d)],$$

$\delta(c,d)$ and y are the only elements in the range of δ that can have $\zeta(t) > t$ ($\zeta(t) = t$ for all $t \leq z$). Hence either $\delta(\alpha,\gamma) = \delta(c,d)$ or $\delta(\alpha,\gamma) = y$ as δ is a strict metric but δ_ζ is not.

Assume $\delta(\alpha,\gamma) = \delta(c,d)$. So

$$\zeta(\delta(\alpha,\beta)) + \zeta(\delta(\beta,\gamma)) \leq \zeta(\delta(c,d)).$$

Without loss of generality the assumption $\delta(\alpha,\beta) = x$ may be made as one of the two left hand terms must come from the partition $\{x,y\}$ and $y > \delta(c,d)$. Furthermore as it is not the case that $[x,y] <_c [\delta(a,b),\delta(c,d)]$ (the transitivity of $<_c$ would lead to a contradiction of the premise), the following holds

$$\delta(a,b) - x < y - \delta(c,d).$$

Combining inequalities gives

$$\zeta(x) + \zeta(\delta(\beta,\gamma)) + \delta(a,b) - x < y - \delta(c,d) + \zeta(\delta(c,d)).$$

But by the convexity of ζ and $y > \delta(c,d)$ the inequality $\zeta(y) - y > \zeta(\delta(c,d)) - \delta(c,d)$ holds. Hence by the definition of ζ

$$\zeta(\delta(\beta,\gamma)) + \zeta(\delta(a,b)) < \zeta(y).$$

This shows that if the triangle inequality does not hold for δ_ζ, then there exists some $\alpha,\beta,\gamma \in X$ such that

$$\zeta(\delta(\alpha,\beta)) + \zeta(\delta(\beta,\gamma)) \leq \zeta(\delta(\alpha,\gamma))$$

and $\delta(\alpha,\gamma) = y$. We are therefore justified proceeding under the assumption $\delta(\alpha,\gamma) = y$.

Let $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ be a strictly-increasing convex function such that $\delta_f = f(\delta)$ is a dissimilarity map (i.e. $f(0) = 0$) and $f(x) + f(y) = f(w) + f(z)$. Note if f is scaled by a positive number, it is still convex (by Proposition 3.7) and if the four-point condition holds for δ_f, then it still holds for the scaled f. Therefore without loss of generality the

assumption $f(y) = \zeta(y)$ (else scale f so it has this property) can be made. But δ_ζ is not a strict metric so there exists some $\alpha, \beta, \gamma \in X$ such that $\zeta(\delta(\alpha,\beta)) + \zeta(\delta(\beta,\gamma)) \leq \zeta(\delta(\alpha,\gamma))$ and $\delta(\alpha,\gamma) = y$. Furthermore, as

$$\zeta(\delta(\alpha,\beta)) + \zeta(\delta(\beta,\gamma)) \leq \zeta(y) = f(y) = f(\delta(\alpha,\gamma))$$

and by Proposition 4.23, the inequality

$$f(\delta(\alpha,\beta)) + f(\delta(\beta,\gamma)) \leq \zeta(y) = f(y) = f(\delta(\alpha,\gamma))$$

holds. Note this relies on the $\{w,z\} \cap \{\delta(\alpha,\beta), \delta(\beta,\gamma)\}$ being non-empty, which holds as $[w,z]$ is a pairwise partition of X ($\delta(\alpha,\beta)$, $\delta(\beta,\gamma)$ and $\delta(\alpha,\gamma) = y$ are all elements of different pairwise partitions of X). Hence δ_f is not a strict metric and $\mathbf{T}_\delta^{c.r.} = \emptyset$. □

Example 5.12. Let $X = \{a,b,c,d\}$ and

$\delta =$

	a	b	c	d
a	0	4	6	8
b	4	0	9	10
c	6	9	0	5
d	8	10	5	0

be a dissimilarity map on X. The pairwise partitions of X, referred to in the rest of this example as the intervals, are given in Figure 5.4a. One can readily verify that δ is a strict metric. A quick (but informal) way to do this is by checking that the larger number in each interval is smaller than the smaller numbers in both other intervals added together. Note, failing this check does not prove δ is not a strict metric. But, regardless, δ does pass this check and hence is a strict metric. Furthermore, $[\delta(a,b), \delta(c,d)] <_c [\delta(a,c), \delta(b,d)]$ holds and $[\delta(a,c), \delta(b,d)]$ and $[\delta(a,d), \delta(b,c)]'$ are $<_c$-incomparable, so to apply Proposition 5.11 we must consider the clue function $\zeta : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ on $\{\delta(a,c), \delta(a,d), \delta(b,c), \delta(b,d)\}$. Explicitly, the clue function ζ is given by

$\mathbb{R}^+ \cup \{0\}$	0	6	8	9	10
$\zeta(\mathbb{R}^+ \cup \{0\})$	0	6	8	9	11

.

Hence the dissimilarity map $\zeta(\delta)$ is given by

	a	b	c	d
a	0	4	6	8
b	4	0	9	11
c	6	9	0	5
d	8	11	5	0

,

and is also a strict metric (one can use Figure 5.4b to show it passes the informal test described earlier in this example). Hence by Proposition 5.11, $\mathbf{T}_{\delta}^{c.r.} = \{ab|cd\}$.

(a) The intervals $[4, 5]$ (in red), $[6, 10]$ (in blue) and $[8, 9]$ (in green).

(b) The intervals $[4, 5]$ (in red), $[6, 11]$ (in blue) and $[8, 9]$ (in green).

Figure 5.4: Intervals associated with Example 5.12.

5.4 Ordinal and convex differences: an example

Example 5.13. Let $X = \{a, b, c, d\}$. Consider the dissimilarity map

$\delta =$

	a	b	c	d
a	0	9	12	15
b	9	0	17	18
c	12	17	0	22
d	15	18	22	0

from Example 3.30. Figure 5.5 shows the pairwise partitions of X for δ. When we previously dealt with this example, weighted phylogenetic X-trees (depicted in Figure 3.8) were explicitly constructed to show $\{ab|cd, ac|bd, ad|bc, abcd\} = \mathbf{T}_{\delta}^{o.e.}$ and $ac|bd \in \mathbf{T}_{\delta}^{c.r.}$. Equipped with the tools from this chapter, we can show a stronger result without resorting to searching for examples. As each pairwise partition of X is distinct and $<_o$-incomparable, by Proposition 5.6, $\{ab|cd, ac|bd, ad|bc, abcd\} = \mathbf{T}_{\delta}^{o.e.}$. Using the method from Example 5.12, one can verify δ is a strict metric. Let

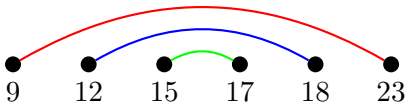


Figure 5.5: The interval $[9, 22]$ (in red) corresponds to the interval $[\delta(a, b), \delta(c, d)]$. Similarly, the interval $[12, 18]$ (in blue) corresponds to $[\delta(a, c), \delta(b, d)]$ and $[15, 17]$ (in green) corresponds to $[\delta(a, d), \delta(b, c)]$.

$\zeta : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ be a clue function on $\{\delta(a, c), \delta(a, d), \delta(b, c), \delta(b, d)\}$. The dissimilarity map $\zeta(\delta)$ is given by

$$\delta = \begin{array}{c|cccc} & a & b & c & d \\ \hline a & 0 & 9 & 12 & 15 \\ b & 9 & 0 & 17 & 18 \\ c & 12 & 17 & 0 & 23 \\ d & 15 & 18 & 23 & 0 \end{array}.$$

Again using the method from Example 5.12, one can verify $\zeta(\delta)$ is a strict metric. Furthermore, the interval relation $[\delta(a, c), \delta(b, d)] <_c [\delta(a, b), \delta(c, d)]$ holds and the intervals $[\delta(a, b), \delta(c, d)]$ and $[\delta(a, d), \delta(b, c)]$ are $<_c$-incomparable. Therefore, by Proposition 5.11, $\{ac|bd\} = \mathbf{T}_\delta^{c.r.}$. This confirms that $\mathbf{T}_\delta^{c.r.}$ can be a non-trivial proper subset of $\mathbf{T}_\delta^{o.e.}$. Furthermore, it highlights the potential advantages of classification algorithms based on the convex assumption over those based on the ordinal assumption.

5.5 Ultrametrics and ordinal equivalence

Definition 5.14. Let δ be a dissimilarity map on X. For every distinct $a, b, c \in X$, if the two greatest values of $\delta(a, b)$, $\delta(b, c)$ and $\delta(a, c)$ are equal, then δ is a *ultrametric*.

Proposition 5.15. *An ultrametric δ is a tree metric.*

Proof. In light of Theorem 3.21, it suffices to show that the ultrametric δ satisfies the four-point condition. □

Definition 5.16. Let $\mathcal{T} = (T, \phi)$ be a weighted X-tree with weight function $w : E(T) \to \mathbb{R}^+$ and $a, b \in X$. An edge $\{u, v\} \in E(T)$ contains the *midpoint*

of a and b if

$$d_{(T,w)}(\phi(a), u) \leq d_{(T,w)}(\phi(b), u),$$

$$d_{(T,w)}(\phi(b), v) \leq d_{(T,w)}(\phi(a), v)$$

and $\{u, v\}$ is on the path connecting $\phi(a)$ and $\phi(b)$. The weighted X-tree $\mathcal{T}$ is *midpoint complete* if each of the edges in $E(T)$ contains a midpoint.

Lemma 5.17. *Let $\mathcal{T} = (T, \phi)$ be a weighted phylogenetic X-tree with weight function $w : E(T) \to \mathbb{R}^+$. Let $a \in X$ and $u \in E(T)$ such that $\{\phi(a), u\}$ is a minimal weight pendant edge. If $\mathcal{T}$ is midpoint complete, then there exists an element $b \in X$ such that $\{\phi(b), u\}$ is a pendant edge with $w(\{\phi(a), u\}) = w(\{\phi(b), u\})$.*

Proof. The minimal weight pendant edge $\{\phi(a), u\}$ must contain the midpoint of a and some $b \in X$. Hence, the relation $d_{(T,w)}(\phi(b), u) \leq w(\{\phi(a), u\})$ holds. As $\{\phi(a), u\}$ is a minimal weighted pendant edge, the relation $w(\{\phi(a), u\}) \leq d_{(T,w)}(\phi(b), u)$ also holds. Combining relations gives $d_{(T,w)}(\phi(b), u) = w(\{\phi(a), u\})$. Lastly, $\{\phi(b), u\}$ must be a pendant edge in $E(T)$ as the path from $\phi(b)$ to u has the same weight (given by $d_{(T,w)}(\phi(b), u)$) as the minimal weight pendant edge $\{\phi(a), u\}$. □

Proposition 5.18. *Let $\mathcal{T} = (T, \phi)$ be a binary weighted phylogenetic X-tree with weight function $w : E(T) \to \mathbb{R}^+$. If $\mathcal{T}$ is midpoint complete, then $d_{(\mathcal{T},w)}$ is an ultrametric on X.*

Proof. This proof is by induction on the size of X. Assume $|X| = 3$. If $\mathcal{T}$ is midpoint complete, then by Lemma 5.17 there exist $a, b \in X$ and $u \in E(T)$ such that $\{\phi(b), u\}$ and $\{\phi(a), u\}$ are minimal weight pendant edges. Let c be the remaining distinct element in X. As $d_{(T,w)}(\phi(b), u) = d_{(T,w)}(\phi(a), u)$ and $d_{(T,w)}(\phi(a), u) \leq d_{(T,w)}(\phi(c), u)$,

$$d_{(\mathcal{T},w)}(a, b) \leq d_{(\mathcal{T},w)}(a, c) = d_{(\mathcal{T},w)}(b, c).$$

Hence, $d_{(\mathcal{T},w)}$ is an ultrametric.

Suppose the result yields when $|X| = k$ for some $k \in \{n \in \mathbb{Z} : n \geq 3\}$. Consider X such that $|X| = k + 1$. By Lemma 5.17, there exist $a, b \in X$

and $u \in E(T)$ such that $\{\phi(b), u\}$ and $\{\phi(a), u\}$ are minimal weight pendant edges. As $\mathcal{T}$ is binary and u is an internal vertex, u is adjacent to one remaining distinct vertex $v \in E(T)$ (i.e. $v \neq \phi(a)$ and $v \neq \phi(b)$). By removing the vertices $\phi(b)$ and u from $V(T)$ and their associated edge from $E(T)$, a new binary weighted binary phylogenetic X'-tree $\mathcal{T}' = (T', w')$ is obtained. It is implied that the set of edges $E(T')$ is updated with the new edge $\{\phi(a), v\}$ and that $w'(\{\phi(a), v\}) = d_{(T,w)}(\phi(a), v)$. As $\mathcal{T}$ was midpoint complete, $\mathcal{T}'$ is also midpoint complete. This can be made explicit by considering each edge in T' as containing the same midpoint as is contained in the corresponding edge in T. This maintains midpoint completeness as any midpoint of b and some other vertex $c \in X'$ is also the midpoint of a and c. By the induction hypothesis and as $\mathcal{T}'$ is midpoint complete with $|X'| = k$, the dissimilarity map $d_{(\mathcal{T}',w)}$ is an ultrametric. For $d_{(\mathcal{T},w)}$ to be an ultrametric, all triples (i.e. an element of the set $\{A \subset X : |A| = 3\}$) must satisfy the ultrametric condition. Triples without b in hold the condition trivially as $d_{(\mathcal{T}',w)}$ is ultrametric. Triples with b but not a in also satisfy the ultrametric condition as for all $c \in X$
$\{a, b\}$, $d_{(\mathcal{T}',w)}(a, c) = d_{(\mathcal{T},w)}(b, c)$ and $d_{(\mathcal{T}',w)}$ is ultrametric. The remaining case of triples containing both a and b is the same as the base case as $\{\phi(b), u\}$ and $\{\phi(a), u\}$ are both minimal weight pendant edges. Hence, $d_{(\mathcal{T},w)}$ is an ultrametric. □

This final example shows that just because a tree metric δ with representation $\mathcal{T}$ has the property $\mathbf{T}_{\delta}^{o.e.} = \{\mathcal{T}\}$, it does not necessarily mean δ is an ultrametric.

Example 5.19. Let $X = \{a, b, c, d\}$. Consider the strict metric

$$\delta = \begin{array}{c|cccc} & a & b & c & d \\ \hline a & 0 & 4 & 8 & 9 \\ b & 4 & 0 & 10 & 11 \\ c & 8 & 10 & 0 & 7 \\ d & 9 & 11 & 7 & 0 \end{array}.$$

As δ satisfies the four-point condition and has type $ab|cd$ it has the tree metric representation $ab|cd$ (by Proposition 3.20). Furthermore, using Propo-

sition 5.4, one can show $\mathbf{T}_{\delta}^{o.e.} = \{ab|cd\}$. But δ is not an ultrametric as $\delta(a,b) < \delta(a,c) < \delta(b,c)$.

Bibliography

H.-J. Bandelt and A. W. Dress. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Molecular phylogenetics and evolution*, 1(3):242–252, 1992.

J. A. Bondy and R. L. Hemminger. Graph reconstructiona survey. *Journal of Graph Theory*, 1(3):227–268, 1977.

F. Bonnot, A. Guénoche, and X. Perrier. Properties of an order distance associated with a tree distance. In *Ordinal and Symbolic Data Analysis*, pages 252–261. Springer, 1996.

O. P. Buneman. The recovery of trees from measures of dissimilarity. *Mathematics in the archaeological and historical sciences*, 1971.

C. R. Darwin. Notebook b: Transmutation, 1837.

P. Edman and G. Begg. A protein sequenator. *European Journal of Biochemistry*, 1(1):80–91, 1967.

J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981.

W. M. Fitch. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology*, 20(4):406–416, 1971.

A. Guénoche. Ordinal properties of tree distances. *Discrete mathematics*, 192(1):103–117, 1998.

T. H. Jukes and C. R. Cantor. Evolution of protein molecules. *Manmmalian Protein Metabolism*, pages 21–132, 1969.

S. Kannan and T. Warnow. *Tree reconstruction from partial orders*. Springer, 1993.

P. E. Kearney. A six-point condition for ordinal matrices. *Journal of Computational Biology*, 4(2):143–156, 1997.

P. E. Kearney. The ordinal quartet method. In *Proceedings of the second annual international conference on Computational molecular biology*, pages 125–134. ACM, 1998.

P. E. Kearney, R. Hayward, and H. Meijer. Evolutionary trees and ordinal assertions. *Algorithmica*, 25(2-3):196–221, 1999.

N. Robertson and P. D. Seymour. Graph minorsa survey. *Surveys in combinatorics*, 103:153–171, 1985.

U. W. Roshan, T. Warnow, B. M. E. Moret, and T. L. Williams. Rec-i-dcm3: a fast algorithmic technique for reconstructing phylogenetic trees. In *Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceedings. 2004 IEEE*, pages 98–109. IEEE, 2004.

S. B. Russ. A translation of Bolzano's paper on the intermediate value theorem. *Historia Mathematica*, 7(2):156–185, 1980.

N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4): 406–425, 1987.

F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977.

C. Semple and M. A. Steel. *Phylogenetics*, volume 24. Oxford University Press on Demand, 2003.

S. Snir, T. Warnow, and S. Rao. Short quartet puzzling: A new quartet-based phylogeny reconstruction algorithm. *Journal of Computational Biology*, 15(1):91–103, 2008.

R. R. Sokal and P. H. Sneath. Numerical taxonomy. *Freemont, San Francisco, CA*, 1963.

K. Strimmer and A. Von Haeseler. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution*, 13(7):964–969, 1996.

K. Tamura. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and g+ c-content biases. *Molecular Biology and Evolution*, 9(4):678–687, 1992.

K. A. Zaretskii. Constructing a tree on the basis of a set of distances between the hanging vertices. *Uspekhi Matematicheskikh Nauk*, 20(6):90–92, 1965.