On the Approximability of Numerical Taxonomy

(Fitting Distances by Tree Metrics)*

Richa Agarwala† DIMACS

Vineet Bafna‡ DIMACS

Martin Farach§ Rutgers University

Babu Narayanan¶ DIMACS

Mike Paterson‖ University of Warwick

Mikkel Thorup** University of Copenhagen

December 12, 1995

Abstract

We consider the problem of fitting an $n \times n$ distance matrix D by a tree metric T. Let ε be the distance to the closest tree metric under the L_∞ norm, that is, $\varepsilon = \min_T\{\| T - D \|_\infty\}$. First we present an $O(n^2)$ algorithm for finding an additive tree T such that $\| T - D \|_\infty \leq 3\varepsilon$. Second we show that it is $\mathcal{NP}$-hard to find a tree T such that $\| T - D \|_\infty < \frac{9}{8}\varepsilon$. This paper presents the first algorithm for this problem with a performance guarantee.

*A preliminary version of this paper appeared in SODA '96

†DIMACS, Rutgers University, Piscataway, NJ 08855, USA. (*agarwala@dimacs.rutgers.edu*) Supported by Special Year National Science Foundation grant BIR-9412594.

‡(*bafna@dimacs.rutgers.edu*) Supported by Special Year National Science Foundation grant BIR-9412594.

§Department of Computer Science, Rutgers University, Piscataway, NJ 08855, USA. (*farach@cs.rutgers.edu, http://www.cs.rutgers.edu/~farach*) Supported by NSF Career Development Award CCR-9501942.

¶(*bon@dimacs.rutgers.edu*) Supported by a DIMACS postdoctoral fellowship under grants STC-88-09648 and 91-19999.

‖Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK. (*msp@dcs.warwick.ac.uk*) Supported in part by the ESPRIT Basic Research Action Programme of the EC under contract No. 7141 (project ALCOM II).

**Department of Computer Science, University of Copenhagen, Universitetsparken 1, 2100 Kbh. Ø, Denmark. (*mthorup@diku.dk, http://www.diku.dk/~mthorup*). This work was done while visiting DIMACS.

1 Introduction

One of the most common methods for clustering numeric data involves fitting the data to a *tree metric*, which is defined by a weighted tree spanning the points of the metric, the distance between two points being the sum of the weights of the edges of the path between them. Not surprisingly, this problem, the so-called *Numerical Taxonomy* problem, has received a great deal of attention (see [2, 7, 8] for extensive surveys) with work dating as far back as the beginning of the century [1]. Fitting distances by trees is an important problem in many areas. For example, in statistics, the problem of clustering data into hierarchies is exactly the tree fitting problem. In "historical sciences" such as paleontology, historical linguistics, and evolutionary biology, tree metrics represent the branching processes which lead to some observed distribution of data. Thus, the numerical taxonomy problem has been, and continues to be, the subject of intense research.

In particular, consider the case of evolutionary biology. By comparing the DNA sequences of pairs of species, biologists get an estimate of the evolutionary time which has elapsed since the species separated by a speciation event. A table of pairwise distances is thus constructed. The problem is then to reconstruct the underlying evolutionary tree. Dozens of heuristics for this problem appear in the literature every year (see, e.g., [8]).

The numerical taxonomy problem is usually cast in the following terms. Let S be the set of species under consideration.

The Numerical Taxonomy Problem

Input: $D : S^2 \rightarrow \Re_{\geq 0}$, a distance matrix.

Output: A *tree metric* T which spans S and *fits* D.

This definition leaves two points unanswered: first, what kind of tree metric, and second, what does it mean for a metric to fit D? Typically we are talking about any tree metric, but sometimes we want to restrict ourselves to *ultrametrics* defined by rooted trees where the distance to the root is the same for all points in S. In order to distinguish specific types of tree metrics, such as ultrametrics, from the general case, we will refer to unrestricted tree metrics as *additive*. There may be no tree metric T coinciding exactly with D, so by "fitting" we mean approximating D under norms such as L_1, L_2, or L_∞. That is, for $k = 1, 2, \ldots, \infty$, we want to find a tree metric T minimizing $\| T - D \|_k$.

History The numerical taxonomy problem for additive metric fitting under L_k norms was explicitly stated in its current form in 1967 [4]. Since then it has collected an extensive literature. In 1977 [10], it was shown that if there is a tree T coinciding exactly with D, it is unique and constructible in linear, i.e., $O(|S|^2)$, time. Unfortunately there is typically no tree T coinciding exactly with D, and in 1987 [5], it was shown that for L_1 and L_2, the numerical taxonomy problem is $\mathcal{NP}$-hard, both in the additive and in the ultrametric cases. Additional complexity results appear in [9].

The only positive fitting result is from 1993 [6] and shows that under the L_∞ norm an optimal ultrametric is polynomially computable, in fact in linear time. However, while ultrametrics have interesting special case applications, the fundamental problem in the area of numerical taxonomy is that of fitting D by general tree metrics. Unfortunately no provably good algorithms existed for fitting distances by additive metrics, and in [6] the Numerical Taxonomy Problem for general tree metrics under the L_∞ norm was posed as a major open problem.

Our Results We consider the Numerical Taxonomy Problem for additive metrics under the L_∞ norm as suggested in [6]. Let ε be the distance to the closest tree metric under the L_∞ norm, that is, $\varepsilon = \min_T\{\| T - D \|_\infty\}$. First we present an $O(n^2)$ algorithm for finding an additive tree T such that $\| T - D \|_\infty \leq 3\varepsilon$. We complement this result not only by finding that an L_∞-optimal solution is $\mathcal{NP}$-hard, but we also rule out arbitrarily close approximations by showing that it is $\mathcal{NP}$-hard to find a tree T such that $\| T - D \|_\infty < \frac{9}{8}\varepsilon$.

Our algorithm is achieved by transforming the general tree metric problem to that of ultrametrics with a loss of a factor of 3 on the approximation ratio. Since the ultrametric problem is optimally solvable, our first result follows. We also generalize our transformation from the general tree metric to ultrametrics under any L_k norm.

The paper is organized as follows. After some preliminary definitions in Section 2, we give our 3-approximation algorithm in Section 3. We show in Section 4 that our analysis is tight, and that some natural "improved" heuristics do not help in the worst case. In Section 5, we give our $\mathcal{NP}$-completeness and non-approximability proofs. Finally, in Section 6, we generalize our reduction from L_∞ to norms with finite k.

2 Preliminaries

We present some basic definition.

Definition 2.1 *A* metric *on a set $S = \{1, \ldots, n\}$ is a function $D : S^2 \to \Re_{\geq 0}$ such that*

- $D[x,y] = 0 \iff x = y$,
- $D[x,y] = D[y,x]$,
- $D[x,y] \leq D[x,z] + D[z,y]$ (the triangle inequality).

Likewise, $D : S^2 \to \Re_{\geq 0}$ is a *quasi-metric* if it satisfies the first two conditions. For (quasi-)metrics A and B, $A + B$ is the usual matrix addition, i.e., $(A+B)[i,j] = A[i,j] + B[i,j]$.

Definition 2.2 *A (quasi-)metric D is* (quasi-)additive *if, for all points a, b, c, d,*

$$D[a,b] + D[c,d] \leq \max\{D[a,c] + D[b,d], D[a,d] + D[b,c]\}.$$

This inequality is known as the *4-point condition.*

Theorem 2.3 (Buneman [3]) *A metric is additive if and only if it is a tree metric.*

Definition 2.4 *A metric D is an* ultrametric *if, for all points a, b, c,*

$$D[a,b] \leq \max\{D[a,c], D[b,c]\}.$$

As noted above, an ultrametric is a type of tree metric.

Definition 2.5 *A quasi-metric D is a* centroid *quasi-metric if* $\exists l_1, \ldots, l_n$ *such that* $\forall i \neq j$, $D[i,j] = l_i + l_j$.

A centroid quasi-metric D is a *centroid metric* if $l_i \geq 0$ for all i. A centroid metric is a type of tree metric since it can be realized by a weighted tree with a star topology and edge weights l_i.

The k-norms are formally defined as follows.

Definition 2.6 *For $n \times n$ real-valued matrices M and $k \geq 1$, define the k-norm, sometimes denoted L_k, by*

$$\| M \|_k = (\sum_{i<j} | M[i,j] |^k)^{\frac{1}{k}},$$

$$\| M \|_\infty = \max_{i<j}\{| M[i,j] |\}.$$

We define the *Additive*$_k$ problem as, given a matrix D, output an additive metric A minimizing $\| D - A \|_k$. Similarly, the *Ultrametric*$_k$ problem is, given a matrix D, output an ultrametric U minimizing $\| D - U \|_k$.

3 Upper Bound

Let $m_a = \max_i\{D[a,i]\}$. Let C^a be the centroid metric with $l_i = m_a - D[a,i]$, i.e., $C^a[i,j] = l_i + l_j = 2m_a - D[a,i] - D[a,j]$.

Lemma 3.1 ([2, Th.3.2]) *For all a, D is quasi-additive if and only if $D + C^a$ is ultrametric.*

Lemma 3.2 ([2, Cor.3.3]) *Given an additive metric A and a centroid quasi-metric Q, $A + Q$ is additive if and only if $A + Q$ satisfies the triangle inequality.*

Let D be a distance matrix. We define $\mathcal{A}(D)$ to be (one of) the additive metrics such that $\| D - \mathcal{A}(D) \|_\infty = \min_A \| D - A \|_\infty$ for all additive metrics A. For point a, we say a metric M is a-restricted if $\forall i, M[a,i] = D[a,i]$. We define $\mathcal{A}^a(D)$ to be (one of) the a-restricted additive metrics such that $\| D - \mathcal{A}^a(D) \|_\infty = \min_A \| D - A \|_\infty$ for all a-restricted additive metrics A. In other words, $\mathcal{A}^a(D)$ is an optimal a-restricted additive tree for D. We will sometimes refer to such a tree as *a-optimal*. Similarly, we define $\mathcal{U}(D)$ to be an optimal ultrametric tree for D. Note that the functions, $\mathcal{A}(), \mathcal{A}^a()$, and $\mathcal{U}()$ need not be single-valued. In the following, we will let the output be an arbitrary optimal metric, unless otherwise noted. Recall that $\mathcal{U}()$ is computable in $O(n^2)$ time [6].

Lemma 3.1 suggests that we may be able to approximate the closest additive metric to D by approximating the closest ultrametric to $D + C^a$, i.e., by computing $\mathcal{U}(D + C^a) - C^a$, for some point a. Lemma 3.2 tells us that we need to guarantee the triangle inequality for the final metric to show that it is additive. Thus we need to modify our heuristic. Specifically, for any point a, we will show that $\| D - \mathcal{A}^a(D) \|_\infty \leq 3\| D - \mathcal{A}(D) \|_\infty$, and we will give a modification $\mathcal{U}^a()$ of $\mathcal{U}()$ such that $\mathcal{A}^a(D) = \mathcal{U}^a(D + C^a) - C^a$. We will use the following result implicit in FKW [6]

Theorem 3.3 *Consider a graph $G = (V, E)$ along with weight functions L, M on the edges and a real value h such that $L[i,j] \leq M \leq h$ for all i, j. There is an $O(n^2)$ algorithm to compute an ultrametric tree U in which for all i, j, $L[i,j] \leq U[i,j] \leq h$, and $\| M - U \|_\infty$ is minimized.*

Proof: Identical to the proof of Theorem 5 in FKW [6]. Note that the upper and lower bounds on the values are now given by $\min\{M[i,j] + \varepsilon, h\}$ and $\max\{M[i,j] - \varepsilon, L[i,j]\}$, respectively. Further, for all ε, a minimum spanning tree for G with edge weight $M[i,j]$ on edge (i,j) is also a minimum spanning tree for G with edge weight $\min\{M[i,j] + \varepsilon, h\}$ on the edge (i,j). ∎

Corollary 3.4 *Consider a graph $G = (V, E)$ along with weight functions L, M on the edges and a real value h such that $L[i,j] \leq h$ for all i, j (M can be arbitrary). There is an $O(n^2)$ algorithm to compute an ultrametric tree U in which for all i, j, $L[i,j] \leq U[i,j] \leq h$, and $\| M - U \|_\infty$ is minimized.*

Proof: Compute a matrix of distances M' given by

$$M'[i,j] = \begin{cases} M[i,j] & \text{if } L[i,j] \leq M[i,j] \leq h \\ h & \text{if } M[i,j] > h \\ L[i,j] & \text{if } M[i,j] < L[i,j] \end{cases}$$

and compute a minimum spanning tree for G with edge weights given by M'. Now follow the outline in Theorem 3.3, using this MST, but computing deviation ε from M. ∎

3.1 The L_∞ Approximation

The *stem* of a leaf is the edge incident on it.

Lemma 3.5 *For all points a, $\| D - \mathcal{A}^a(D) \|_\infty \leq 3\| D - \mathcal{A}(D) \|_\infty$.*

Proof: For all i, j, let $\varepsilon[i,j] = \mathcal{A}(D)[i,j] - D[i,j]$, and $\varepsilon = \max_{i,j}\{|\varepsilon[i,j]|\}$.
Derive an a-restricted tree $T^{/a}$ from $\mathcal{A}(D)$ as follows. If a point i needs to be moved further away from a, that is, if $\mathcal{A}(D)[a,i] - D[a,i]$ is negative, we simply increase the length of its stem. To move i closer (when $\mathcal{A}(D)[a,i] - D[a,i]$ is positive), if the stem is too short, we might have to let i pass some interior vertices in the obvious way. In either case, no point i is moved more than $|\varepsilon[a,i]|$. Now, $T^{/a}$ is additive by construction, and for all i, $T^{/a}[a,i] = D[a,i]$. Further, for all i, j,

$$\begin{aligned} |D[i,j] - T^{/a}[i,j]| &\leq |\mathcal{A}(D)[i,j] - T^{/a}[i,j]| + |D[i,j] - \mathcal{A}(D)[i,j]|, \\ &\leq (|\varepsilon[a,i]| + |\varepsilon[a,j]|) + |\varepsilon[i,j]| \\ &\leq 3\varepsilon. \end{aligned}$$

Finally, by the optimality of $\mathcal{A}^a(D)$,

$$\| D - \mathcal{A}^a(D) \|_\infty \leq \| D - T^{/a} \|_\infty \leq 3\varepsilon.$$

∎

Lemma 3.6 *$\mathcal{A}^a(D)$ can be computed in polynomial time.*

Proof: We say an ultrametric U is a-restricted (with respect to D) if it satisfies the following constraints:

$$
\begin{aligned}
2m_a \geq \ U[i,j] &\geq \ 2\max\{l_i, l_j\}, \text{ for all } i,j, && (1)\\
U[a,i] &= \ 2m_a, \text{ for all } i \neq a. && (2)
\end{aligned}
$$

For any distance matrix M, define $\mathcal{U}^a(M)$ to be an a-restricted ultrametric minimizing $\| M - \mathcal{U}^a(M) \|_\infty$. Note that for all i,j, $\mathcal{U}^a(M)[i,j] \leq 2m_a$. Therefore, by corollary 3.4, $\| M - \mathcal{U}^a(M) \|_\infty$ can be computed in $O(n^2)$ time.

Let $T = \mathcal{U}^a(D + C^a) - C^a$. We now show that $T = \mathcal{A}^a(D)$.

CLAIM 3.6A *T is an a-restricted additive metric.*

PROOF: Let $D^a = D + C^a$. Constraint (2) implies that T is a-restricted. By Lemma 3.2, we only need to show that T satisfies the triangle inequality, i.e.,

$$
\begin{aligned}
& T[i,j] &\leq\ & T[i,k] + T[k,j], \text{ for all distinct } i,j,k\\
\Leftrightarrow\ & \mathcal{U}^a(D^a)[i,j] - C^a[i,j] &\leq\ & \mathcal{U}^a(D^a)[i,k] - C^a[i,k] + \mathcal{U}^a(D^a)[k,j] - C^a[k,j]\\
\Leftrightarrow\ & \mathcal{U}^a(D^a)[i,j] &\leq\ & \mathcal{U}^a(D^a)[i,k] + \mathcal{U}^a(D^a)[k,j] - 2l_k\\
\Leftrightarrow\ & \mathcal{U}^a(D^a)[i,j] &\leq\ & \max\{\mathcal{U}^a(D^a)[i,k], \mathcal{U}^a(D^a)[k,j]\}\\
& & & + \min\{\mathcal{U}^a(D^a)[i,k], \mathcal{U}^a(D^a)[k,j]\} - 2l_k.
\end{aligned}
$$

Now, since $\mathcal{U}^a(D^a)$ is ultrametric,

$$\mathcal{U}^a(D^a)[i,j] \leq \max\{\mathcal{U}^a(D^a)[i,k], \mathcal{U}^a(D^a)[k,j]\}.$$

Also, $\min\{\mathcal{U}^a(D^a)[i,k], \mathcal{U}^a(D^a)[k,j]\} \geq 2l_k$ by Constraint (1). Hence, the claim is proved. □

CLAIM 3.6B *$\mathcal{A}^a(D) + C^a$ is an a-restricted ultrametric.*

PROOF: From Lemma 3.1, $\mathcal{A}^a(D) + C^a$ is ultrametric. To show that Constraint (2) is satisfied, we note that

$$T'[a,i] = \mathcal{A}^a(D)[a,i] + C^a[a,i] = D[a,i] + l_i + l_a = 2m_a.$$

For Constraint (1), we take $i,j \neq a$, and then

$$
\begin{aligned}
T'[a,j] &\leq\ T'[j,i] + T'[a,i] - 2l_i\\
\Rightarrow\ 2m_a &\leq\ T'[j,i] + 2m_a - 2l_i\\
\Rightarrow T'[j,i] &\geq\ 2l_i.
\end{aligned}
$$

By symmetry, $T'[j,i] \geq 2l_j$. Also,

$$
\begin{aligned}
T'[i,j] &= \mathcal{A}^a(D)[i,j] + l_i + l_j\\
&\leq \mathcal{A}^a(D)[a,i] + \mathcal{A}^a(D)[a,j] + l_i + l_j\\
&= 2m_a
\end{aligned}
$$

Therefore, Constraint (1) is also satisfied and Claim 2 is proved. □

Finally,

$$\begin{aligned}
\| T - D \|_\infty &\geq \| \mathcal{A}^a(D) - D \|_\infty \text{ (by Claim 3.6A)} \\
&= \| (\mathcal{A}^a(D) + C^a) - (D + C^a) \|_\infty \\
&\geq \| \mathcal{U}^a(D + C^a) - (D + C^a) \|_\infty \text{ (by Claim 3.6B)} \\
&= \| T - D \|_\infty \text{ (by construction).}
\end{aligned}$$

Therefore, $\| T - D \|_\infty = \| \mathcal{A}^a(D) - D \|_\infty$. This proves the lemma. ■

Lemmas 3.5 and 3.6 imply:

Theorem 3.7 *Given an $n \times n$ distance matrix D, we can find a tree T in $O(n^2)$ time such that*

$$\| T - D \|_\infty \leq 3\| \mathcal{A}(D) - D \|_\infty.$$

4 Tightness of analysis

Lemma 4.1 *There is an $n \times n$ distance matrix D such that, for all points c,*

$$\frac{\| D - \mathcal{A}^c(D) \|_\infty}{\| D - \mathcal{A}(D) \|_\infty} = 3.$$

Lemma 4.1 states that the constant in Lemma 3.5 is tight, and that it is not improved by trying different values of c.

Proof: First we will prove the lemma for some point c, and later we will generalize the construction to work for all points c.

Consider the following distance matrix D for the points a_1, a_2, b_1, b_2, c: $D[c, a_i] = x + y - \varepsilon/2$, $D[c, b_i] = x + y + \varepsilon/2$, $D[a_i, b_i] = 0$, $D[a_1, a_2] = 2y + 2\varepsilon$, $D[b_1, b_2] = 2y - 2\varepsilon$ $(i = 1, 2)$. Finally, for completion, for $i \neq j$, set $D[a_i, b_j] = 2y$.

Given that $y \gg \varepsilon$, it is easy to see that the two solutions, $\mathcal{A}(D)$ and $\mathcal{A}^c(D)$, in Figure 4.1 are optimal and c-optimal, respectively. Hence $\| D - \mathcal{A}(D) \|_\infty = \varepsilon$ and $\| D - \mathcal{A}^c(D) \|_\infty = 3\varepsilon$. Note that $\mathcal{A}(D)$ is the unique optimal tree. In contrast, for $\mathcal{A}^c(D)$, without violating c-optimality, we could make some variation by giving a_i a small stem of length at most ε. Also note that for any $p \neq c$, $\| D - \mathcal{A}^p(D) \|_\infty = 2\varepsilon$.

In order to get the same result independent of the choice of c, we basically connect three constructions of the above type facing each other such that any point of the one system plays the role of c for the points in one of the other systems. Fix the x in D to 0. We will now make a distance matrix D^* over points of the form (i, p) where $i = 0, 1, 2$ and $p = a_1, a_2, b_1, b_2$. For all $i \in \{0, 1, 2\}$ and $p, q \in \{a_1, a_2, b_1, b_2\}$, set $D^*[(i, p), (i, q)] = D[p, q]$ and set $D^*[(i, p), (i + 1 \bmod 3, q)] = T[c, p] + D[c, q]$. From the considerations concerning D, it follows that we have a tree T^* with error ε, given by $T^*[(i, p), (i, q)] = T[p, q]$, and for $i \neq j$, $T^*[(i, p), (j, q)] = T[c, p] + T[c, q]$. Thus $\| D^* - \mathcal{A}(D^*) \|_\infty \leq \varepsilon$. For any point (i, p), from the previous analysis of D with $x := D[c, p]$ in relation to the points $(i + 1 \bmod 3, q)$, it follows that $\| D^* - \mathcal{A}(D^*)(i, p) \|_\infty \geq 3\varepsilon$. Combined with Lemma 3.5, this gives $\| D^* - \mathcal{A}(D^*) \|_\infty = \varepsilon$ and $\| D^* - \mathcal{A}(D^*)(i, p) \|_\infty = 3\varepsilon$. ■

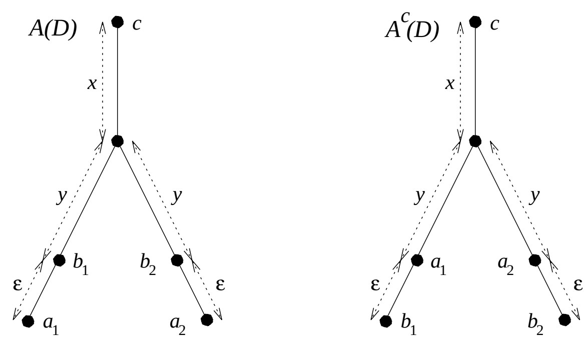

Figure 4.1: Trees approximating D.

Some rather involved examples show that there are c-optimal trees for which changing the edge-lengths cannot bring the error down below $3\varepsilon - o(1)$. Thus there is no significant worst-case advantage to the obvious heuristic of changing the edge-lengths optimally using linear programming.

5 Lower bound

In this section, we show that the problem of finding a tree T such that $\| T - D \|_\infty < \frac{9}{8}\varepsilon$ is $\mathcal{NP}$-hard. First, we show that a decision version of the Numerical Taxonomy Problem is $\mathcal{NP}$-complete.

The Numerical Taxonomy Problem

Input: A distance matrix $D : S^2 \rightarrow \Re_{\geq 0}$, and a threshold $\Delta \in \Re_{\geq 0}$.

Question: Is there a *tree metric* T which spans S and for which $\| T - D \|_\infty \leq \Delta$.

Theorem 5.1 *The Numerical Taxonomy Problem is $\mathcal{NP}$-complete.*

Proof: From Definitions 2.1 and 2.2, the problem is in NP.

We show $\mathcal{NP}$-completeness by reduction from 3SAT. For an instance of 3SAT with variables $x_1, \ldots, x_n$ and clauses $C_1, \ldots, C_k$, we will construct a distance matrix D such that the 3SAT expression is satisfiable if and only if $\| D - \mathcal{A}(D) \|_\infty \leq \Delta = 2$. Let integer r represent some sufficiently large distance (like 10). We construct a distance matrix D to approximate path lengths on a tree with leaves $v, x_i, \bar{x}_i, h_i$ for $1 \leq i \leq n$, and c_j, c'_j, c''_j for $1 \leq j \leq k$. We write $\tilde{x}$ for either x or $\bar{x}$.

To simplify the description of the construction we first present it in the form of a set of inequalities on the distances between the vertices of a tree T, which are expressed later in the required form. For example, we shall write "$T[x_i, \bar{x}_i] \geq 2r$" at first, and realize this constraint eventually by letting $D[x_i, \bar{x}_i] = 2r + 2$. We classify the inequalities as follows.

A: *Literal pairs*

$$T[x_i, \bar{x}_i] \geq 2r, \quad T[\tilde{x}_i, h_i] \leq r, \quad \text{for all } i.$$

These inequalities force h_i to be the midpoint of the path between x_i and $\bar{x}_i$, for all i.
B: *Star-like tree*

$$T[v, \tilde{x}_i] \leq r + 1, \quad T[h_i, h_j] \geq 2, \quad T[h_i, \tilde{x}_j] \geq r, \quad \text{for all } i, j \ (i \neq j).$$

The first inequalities in B, together with those in A, imply $T[v, h_i] \leq 1$ for all i, and we can then use the second inequalities to deduce that $T[v, h_i] = 1$ for all i.

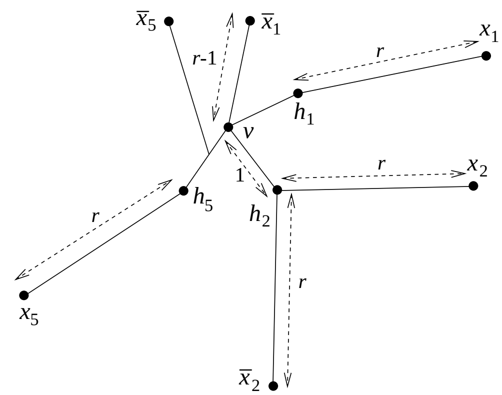

Figure 5.2: Portion of sample layout

The vertex v must be at the center of a star with each h_i at distance 1 from it along separate edges. From each h_i, at least one of the two paths of length r to x_i and $\bar{x}_i$ proceeds away from v. An impression of a general feasible configuration is presented in Figure 5.2.

The essential feature of such configurations, which we shall use in our reduction, is that for each i, *at least one* of x_i and $\bar{x}_i$ is at distance $r + 1$ from v. The final inequalities will represent the satisfaction of clauses by literals. A satisfying literal will correspond to a vertex $\tilde{x}_i$ such that $T[v, \tilde{x}_i] = r - 1$. Clearly, x_i and $\bar{x}_i$ cannot both be satisfying literals.

Now, we present the third set of inequalities that deal with the "clause" vertices c_j, c'_j, c''_j. Specifically, we will show that a clause is satisfied if and only if at least one of its literals $\tilde{y}$ is at a distance less than $r + 1$ from v.
C: *Clause satisfaction*

For each clause $C_j = (y_j, y'_j, y''_j)$ where y_j, y'_j, y''_j are literals, we have three vertices c_j, c'_j, c''_j and the following inequalities (where we drop the subscript for clarity):

$$T[c, y'] \leq r + 1, \ T[c, y''] \leq r + 1,$$

$$T[c', y''] \leq r + 1, \ T[c', y] \leq r + 1,$$

$$T[c'', y] \leq r + 1, \ T[c'', y'] \leq r + 1,$$

$$T[c, c'] \geq 2, \ T[c', c''] \geq 2, \ T[c'', c] \geq 2.$$

If $T[v, y_j]$, $T[v, y'_j]$ and $T[v, y''_j]$ are all $r + 1$, then the first inequalities in C force each of c_j, c'_j, c''_j to coincide with v, contravening the second inequalities.

However, if at least one of these literals is at a distance $r-1$ of v then a configuration of the form illustrated in Figure 5.3 is feasible.

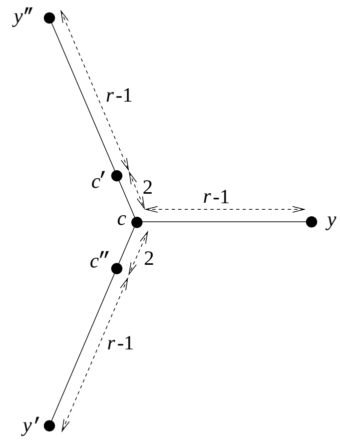

Figure 5.3: Layout of clause vertices

We claim that the complete set of inequalities is satisfiable if and only if the corresponding 3SAT formula is satisfiable. In one direction, suppose that there is a satisfying truth assignment to the logical variables. For each variable, lay out the corresponding tree vertices so that the vertex corresponding to the true literal is at distance $r-1$ from v (the "false" literal will be at distance $r+1$ from v). Each clause has a satisfying literal, therefore, for each j, at least one of y_j, y'_j, y''_j is at distance $r-1$ from v in the tree, thus allowing a legal placement of c_j, c'_j, c''_j. On the other hand, if there is a tree layout satisfying all the inequalities then at least one of y_j, y'_j, y''_j must be within distance $r-1$ of v for each j. Since at most one of x_i and $\bar{x}_i$ can be within $r-1$ of v, the layout yields a (partial) assignment which satisfies the logical formula.

Finally, we construct a distance matrix as follows: for all vertices a, b, and all $x \in \Re_{\geq 0}$, if an inequality is of the form $T[a,b] \geq x$, let $D[a,b] = x+2$. Correspondingly, if the inequality is of the form $T[a,b] \leq x$, let $D[a,b] = x-2$. Observe that any tree metric T' satisfying $\| T' - D \|_\infty \leq \Delta = 2$ must satisfy all the inequalities, and conversely, any tree metric T' satisfying all the inequalities A, B and C must also satisfy $\| T' - D \|_\infty \leq \Delta$. ■

Next, we strengthen Theorem 5.1 to show a hardness-of-approximation result.

Theorem 5.2 *Given a 3SAT instance S, a distance matrix D can be computed in polynomial time such that:*

1. *If S is satisfiable, then $\| D - \mathcal{A}(D) \|_\infty \leq 2$.*
2. *If S is not satisfiable, then $\| D - \mathcal{A}(D) \|_\infty \geq 2 + \frac{1}{4}$.*

Proof: We extend the construction of Theorem 5.1 by relaxing each inequality by a fixed amount δ.

A: *Literal pairs*

$$T[x_i, \bar{x}_i] \geq 2r - \delta, \quad T[\tilde{x}_i, h_i] \leq r + \delta, \quad \text{for all } i.$$

For any three distinct tree vertices (u, v, w), let $\texttt{meet}(u, v, w)$ denote the intersection point of the paths between them. For any variable x_i, we use $\hat{h}_i$ to denote $\texttt{meet}(h_i, x_i, \bar{x}_i)$. From the inequalities in A, it follows that $T[h_i, \hat{h}_i] \leq 3\delta/2$. Also, we have

CLAIM 5.2A *For all $\tilde{x}_i$, $T[\tilde{x}_i, \hat{h}_i] - T[h_i, \hat{h}_i] \geq r - 2\delta$.*

PROOF: By symmetry we may assume that $\tilde{x}_i = x_i$. Then,

$$T[x_i, \hat{h}_i] - T[h_i, \hat{h}_i] = T[x_i, \bar{x}_i] - T[\bar{x}_i, h_i] \geq 2r - \delta - (r + \delta) = r - 2\delta.$$

□

For the second set of inequalities, we provide a node v_i for each variable x_i, instead of having a single node v.
B: *Star-like tree*

$$T v_i \tilde{x}_i \leq r + 1 + \delta, \quad T[v_i, h_i] \geq 1 - \delta, \quad \text{for all } i.$$

$$T[v_i, v_j] \leq 1 + \delta, \quad T[v_i, h_j] \leq 1 + \delta, \quad T[h_i, h_j] \geq 2 - \delta, \quad \text{for all } i, j \ (i \neq j).$$

As in Theorem 5.1, for any variable x, technically we allow both x and $\bar{x}$ to be false. *We interpret $\tilde{x}_i$ as false if* $T[h_i, \texttt{meet}(v_i, h_i, \tilde{x}_i)] \leq T[h_i, \hat{h}_i]$. Note that this allows at most one of x_i and $\bar{x}_i$ to be true.
C: *Clause satisfaction*

For each clause $C = (y_0, y_1, y_2)$ where y_0, y_1, y_2 are literals, we have three vertices c_0, c_1, c_2 and the following inequalities:

$$T[c_0, y_1] \leq r + 1 + \delta, \ T[c_0, y_2] \leq r + 1 + \delta,$$

$$T[c_1, y_2] \leq r + 1 + \delta, \ T[c_1, y_0] \leq r + 1 + \delta,$$

$$T[c_2, y_0] \leq r + 1 + \delta, \ T[c_2, y_1] \leq r + 1 + \delta,$$

$$T[c_0, c_1] \geq 2 - \delta, \ T[c_1, c_2] \geq 2 - \delta, \ T[c_0, c_2] \geq 2 - \delta.$$

Note that if we identify all the nodes v_i, the inequalities are a relaxation of the inequalities in the construction of Theorem 5.1. It follows that if S is satisfiable, then there is a tree T that satisfies these inequalities for all non-negative δ. Consequently, if S is satisfiable, then $\| D - \mathcal{A}(D) \|_\infty \leq 2$.

In the remaining part, we assume that S is not satisfiable. Thus, for any truth assignment there is an unsatisfied clause. Using this unsatisfied clause, we will show that if there is a tree T with $\| D - T \|_\infty \leq 2 + \delta$, then $\delta \geq 1/4$.

By symmetry, we may *restrict our attention to a truth assignment setting all x_i false.* For all i, set $\hat{v}_i = \texttt{meet}(v, h_i, x_i)$, and for each $j \neq i$, set $h_j^i = \texttt{meet}(h_j, h_i, x_i)$. Note that x_i being false means that $T[h_i, \hat{v}_i] \leq T[h_i, \hat{h}_i]$.

CLAIM 5.2B *If $\delta \leq \frac{1}{3}$ then, for any $j \neq i$, $T[h_i, h_j^i] \leq T[h_i, \hat{v}_i]$.*

PROOF: Note that

$$T[\hat{v}_i, v_i] \geq T[h_i, v_i] - T[h_i, \hat{v}_i] \geq 1 - \delta - T[h_i, \hat{h}_i] \geq 1 - \frac{5\delta}{2}. \tag{3}$$

Assume $T[h_i, h_j^i] > T[h_i, \hat{v}_i]$. Then there is a simple path from v_i to $\hat{v}_i$ to h_j^i to h_j. Thus,

$$T[\hat{v}_i, h_j] = [v_i, h_j] - T[v_i, \hat{v}_i] \leq 1 + \delta - (1 - \frac{5\delta}{2}) = \frac{7\delta}{2}.$$

But the inequality $T[h_i, h_j^i] > T[h_i, \hat{v}_i]$ also implies that there is a simple path from h_i to $\hat{v}_i$ to h_j^i to h_j, so

$$\begin{aligned} T[\hat{v}_i, h_j] &= T[h_i, h_j] - T[h_i, \hat{v}_i] \\ &> T[h_i, h_j] - T[h_i, \hat{h}_j] \\ &\geq 2 - \delta - \frac{3\delta}{2} = 2 - \frac{5\delta}{2}. \end{aligned}$$

The two inequalities for $T[\hat{v}_i, h_j]$ imply $\frac{7\delta}{2} \geq 2 - \frac{5\delta}{2}$, or $\delta \geq \frac{1}{3}$. This completes the proof of the lemma. □

CLAIM 5.2C *For all* $i \neq j, \delta \leq \frac{1}{3}$, $T[x_i, x_j] \geq 2r + 2 - 5\delta$.

PROOF: By Claim 5.2B, $T[h_i, h_j^i] \leq T[h_i, v_i]$, and $T[h_i, v_i] \leq T[h_i, \hat{h}] \leq 3\delta/2$. Symmetrically, $T[h_j, h_i^j] \geq 3\delta/2$. Since $T[h_i, h_j] \geq 2 - \delta$ and $\delta \leq 3$, we may conclude that we have a simple path from h_i to h_j^i to h_i^j to h_j, and a simple path from x_i to $\hat{h}_i$ to h_j^i to h_i^j to $\hat{h}_j$ to x_j. Note, however, that $\hat{h}_i$ and h_j^i may coincide, and similarly for h_i^j and $\hat{h}_j$. In conclusion,

$$\begin{aligned} T[x_i, x_j] &= T[x_i, \hat{h}_i] + T[\hat{h}_i, h_j^i] + T[h_j^i, h_i^j] + T[h_i^j, \hat{h}_j] + T[\hat{h}_j, x_j] \\ &\geq T[x_i, \hat{h}_i] + T[h_j^i, h_i^j] + T[\hat{h}_j, x] \\ &= T[x_i, \hat{h}_i] + T[h_i, h_j] - T[h_i, \hat{h}_i] - T[h_j, \hat{h}_j] + T[\hat{h}_j, x_j] \\ &\geq 2(r - 2\delta) + 2 - \delta = 2r + 2 - 5\delta. \end{aligned}$$

For the last derivation, we used Claim 5.2A. □

Finally, we show that if S is not satisfiable, then each tree T that satisfies inequalities A, B, and C must also satisfy $\| D - T \|_\infty \geq 2 + \frac{1}{4}$. If $\delta \geq 1/3$, this is trivially true, so we may assume $\delta < 1/3$. Hence, Claims 5.2B and 5.2C apply.

Let T be a tree that satisfies inequalities A, B, and C with $\| D - T \|_\infty \leq 2 + \delta$. For $i = 0, 1, 2$, let nodes x_i correspond to the three false literals of a clause. Let $p = \mathtt{meet}(x_0, x_1, x_2)$. Then, for $i = 0, 1, 2$,

$$\begin{aligned} T[c_i, p] &= \frac{1}{2}(T[c_i, x_{(i-1)\bmod 3}] + T[c_i, x_{(i+1)\bmod 3}] - T[x_{(i-1)\bmod 3}, x_{(i+1)\bmod 3}]) \\ &\leq \frac{1}{2}\Big((r + 1 + \delta) + (r + 1 + \delta) - (2r + 2 - 5\delta)\Big) \\ &= \frac{7\delta}{2}. \end{aligned}$$

The inequality above follows from inequalities C and Claim 5.2C. Therefore, for $0 \leq i, j \leq 2$, we have $T[c_i, c_j] \leq 7\delta$. The inequalities in C imply $T[c_i, c_j] \geq 2 - \delta$. The combination of these inequalities completes the proof. ∎

Theorem 2 immediately implies a hardness-of-approximation result for the Numerical Taxonomy Problem.

Corollary 5.3 *It is an $\mathcal{NP}$-hard problem, given a distance matrix D, to find an additive metric T such that*

$$\frac{\| D - T \|_\infty}{\| D - \mathcal{A}(D) \|_\infty} < \frac{9}{8}.$$

6 Generalization to Other Norms

First, we show that Lemma 3.5 can be generalized to other norms.

Theorem 6.1 *Let D be a distance matrix, and T be a tree such that $\| D - T \|_p \leq \varepsilon$. Then there exists a point a and an a-restricted tree $T^{/a}$ such that $\| D - T^{/a} \|_p \leq 3\varepsilon$.*

Proof: For any point a, the construction of Lemma 3.5 returns an a-restricted tree $T^{/a}$ such that

$$|T^{/a}[i,j] - D[i,j]| \leq |\varepsilon[i,j]| + |\varepsilon[a,i]| + |\varepsilon[a,j]|, \text{ for all } i, j. \tag{4}$$

Also, by the convexity of the function $|x|^p$ for real x, we have

$$\sum_{i=1}^{k} \frac{|x_i|^p}{k} \geq \left| \frac{\sum_{i=1}^{k} x_i}{k} \right|^p. \tag{5}$$

We continue the proof by an averaging argument. Clearly,

$$\min_a \{(\| T^{/a} - D \|_p)^p\} \leq \frac{\sum_{a=1}^{n} (\| T^{/a} - D \|_p)^p}{n}.$$

We use inequalities (4) and (5) to bound the sum.

$$\begin{aligned}
\sum_{a=1}^{n} (\| T^{/a} - D \|_p)^p &= \sum_{a=1}^{n} \sum_{i=1, i\neq a}^{n} \sum_{j=1, j\neq a}^{n} |\varepsilon[i,j] - \varepsilon[a,i] - \varepsilon[a,j]|^p \\
&\leq 3^{p-1} \sum_{a=1}^{n} \sum_{i=1, i\neq a}^{n} \sum_{j=1, j\neq a}^{n} |\varepsilon[i,j]|^p + |\varepsilon[a,i]|^p + |\varepsilon[a,j]|^p \\
&= 3^p n (\| T - D \|_p)^p.
\end{aligned}$$

The theorem follows. ∎

As in the case of L_∞, we can show that if T is an a-optimal tree for D under L_k, then $T + C^a$ is an optimal a-restricted ultrametric for $D + C^a$ under the same norm. Thus, we conclude with:

Theorem 6.2 *If $A(D)$ is an algorithm which achieves an α-approximation for the a-restricted Ultrametric$_k$ problem and runs in time $T(n^2)$, then there is an algorithm $F(D)$ which achieves a 3α-approximation for the Additive$_k$ problem and runs in $O(nT(n^2))$ time.*

References

[1] R. Baire. *Leçons sur les Fonctions Discontinues.* Paris, 1905.

[2] J-P. Barthélemy and A. Guénoche. *Trees and Proximity Representations.* Wiley, New York, 1991.

[3] P. Buneman. The recovery of trees from measures of dissimilarity. In F. Hodson, D. Kendall and P. Tautu, editors, *Mathematics in the Archaeological and Historical Sciences*, pages 387–395, Edinburgh, 1971. Edinburgh University Press.

[4] L. Cavalli-Sforza and A. Edwards. Phylogenetic analysis models and estimation procedures. *Amer. J. Human Genetics*, 19:233–257, 1967.

[5] W.H.E. Day. Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of Mathematical Biology*, 49(4):461–467, 1987.

[6] M. Farach, S. Kannan, and T. Warnow. A robust model for finding optimal evolutionary trees. *Algorithmica*, 13:155–179 (1995).

[7] P. H. A. Sneath and R. R. Sokal. *Numerical Taxonomy.* W. H. Freeman, San Francisco, California, 1973.

[8] D. L. Swofford and G. J. Olsen. Phylogeny reconstruction. In D. M. Hillis and C. Moritz, editors, *Molecular Systematics*, pages 411–501. Sinauer Associates Inc., Sunderland, MA., 1990.

[9] H.T. Wareham. On the complexity of inferring evolutionary trees. Technical Report #9301, Memorial University of Newfoundland, 1993.

[10] M.S. Waterman, T.F. Smith, M. Singh, and W.A. Beyer. Additive evolutionary trees. *J. Theor. Biol.*, 64:199–213, 1977.