DIMACS Technical Report 95-46
September 1995

On the Approximability of Numerical Taxonomy

(Fitting Distances by Tree Metrics)

by

Richa Agarwala[1] DIMACS

Vineet Bafna[2] DIMACS

Martin Farach[3]. Rutgers University

Babu Narayanan[4] DIMACS

Mike Paterson[5] University of Warwick

Mikkel Thorup[6] University of Copenhagen

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

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

[3]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.

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

[5]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).

[6]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.

DIMACS is a cooperative project of Rutgers University, Princeton University, AT&T Bell Laboratories and Bellcore.

DIMACS is an NSF Science and Technology Center, funded under contract STC–91–19999; and also receives support from the New Jersey Commission on Science and Technology.

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, 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.

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] and [7] for extensive surveys) with work dating as far back as to 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.

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 between 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 talk about 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 metrics fitting under L_k norms was explicitly stated in current form in 1967 [4]. Since then it has collected an extensive literature (for a survey, see [2, 8]). In 1977 [9] 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.

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 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 by not only that finding 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 result follows. We also generalize our transformation from general tree metric to ultrametrics to any L_k norm.

The paper is organized as follows. After some preliminary definitions in Section 2, in Section 3, we give our 3-approximation algorithm. In Section 4, we show 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*

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

For 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 metric D is* 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 in 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. An ultrametric can also be represented by a weighted tree such that $D[i, j]$ is the maximum weight of an edge on the path between i and j.

Definition 2.5 *A metric D is a* centroid *metric if $\exists l_1, \ldots, l_n \geq 0$ such that $\forall i \neq j$, $D[i, j] = l_i + l_j$.*

If any $l_i < 0$, then call D a *centroid quasi-metric.* 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_1, M_2, and $k > 1$, define the k-norm, sometimes denoted L_k, by*

$$\| M_1, M_2 \|_k = (\sum_{i<j} | M_1[i, j] - M_2[i, j] |^k)^{\frac{1}{k}},$$

$$\| M_1, M_2 \|_\infty = \max_{i<j}\{| M_1[i, j] - M_2[i, j] |\}.$$

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]) *D is 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 triangle inequality.*

Let D be a distance matrix. We define $\mathcal{A}(D)$ to be (one of) the additive metrics such that $\| D, \mathcal{A}(D) \|_\infty$ is minimized. 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 $\| D, \mathcal{A}^a(D) \|_\infty$ is minimized. 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 (one of) the ultrametrics such that $\| D, \mathcal{U}(D) \|_\infty$ is minimized. Note that these function, $\mathcal{A}()$, $\mathcal{A}^a()$, 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. Very 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 $\mathcal{A}^a(D) = \mathcal{U}^a(D + C^a) - C^a$.

3.1 The L_∞ Approximation

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

Lemma 3.3 *For all point 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 $T^{/a}[a,i] = D[a,i]$. 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 3.4 *$\mathcal{A}^a(D)$ 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:

$$U[i,j] \geq 2\max\{l_i, l_j\}, \text{ for all } i, j. \tag{1}$$

$$U[a,i] = 2m_a, \text{ for all } i \neq a. \tag{2}$$

For any distance matrix M, define $\mathcal{U}^a(M)$ to be an a-restricted ultrametric minimizing $\| M, \mathcal{U}^a(M) \|_\infty$. $\| M, \mathcal{U}^a(M) \|_\infty$ can be computed in $O(n^2)$ time by a minor modification of FKW [6].

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

Claim 3.4A *T is an a-restricted additive tree.*

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.,

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

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 3.4B *$\mathcal{A}^a(D) + C^a$ 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 $i, j \neq a$,

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

By symmetry, $T'[j,i] \geq 2l_j$. Therefore, Constraint 1 is also satisfied and Claim 2 is proved. □

Finally,

$$\begin{aligned} \| T, D \|_\infty &\geq \| \mathcal{A}^a(D), D \|_\infty \text{ (Claim 3.4A)} \\ &= \| (\mathcal{A}^a(D) + C^a), (D + C^a) \|_\infty \\ &\geq \| \mathcal{U}^a(D + C^a), (D + C^a) \|_\infty \text{ (Claim 3.4B)} \\ &= \| T, D \|_\infty \text{ (by construction).} \end{aligned}$$

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

Lemmas 3.3 and 3.4 imply

Theorem 3.5 *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 3.3 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 in Figure 4.1 are optimal,

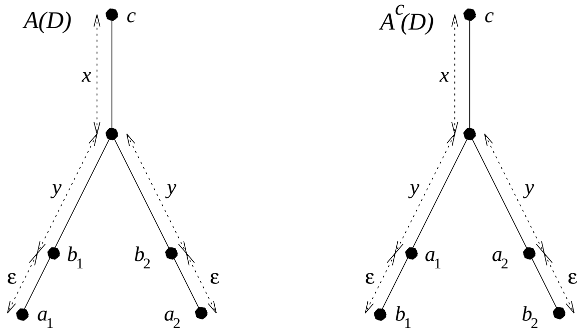

Figure 4.1: Trees approximating D.

respectively c-optimal. Hence $\| D, \mathcal{A}(D) \|_\infty = \varepsilon$ and $\| D, \mathcal{A}^c(D) \|_\infty = 3\varepsilon$. Note that $\mathcal{A}(D)$ is the unique. 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 $\leq \varepsilon$. 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 $E[(i, p), (i, q)] = D[p, q]$ 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 3.3, this gives $\| D^*, \mathcal{A}(D^*) \|_\infty = \varepsilon$ and $\| D^*, \mathcal{A}(D^*)(i, p) \|_\infty = 3\varepsilon$. ∎

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. For the sake of clarity, we only show the hardness of finding ε exactly, and leave the details of the non-approximability result to the full version of the paper.

Theorem 5.1 *It is an $\mathcal{NP}$-complete problem to determine for a given distance matrix D and a threshold Δ if $\| D, \mathcal{A}(D) \|_\infty \leq \Delta$.*

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 $\| D, \mathcal{A}(D) \|_\infty \leq \Delta = 2$. We construct the 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 $\bar{x}$. The integer r represents some sufficiently large distance (like 10) and is used to clarify the construction.

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 may classify the inequalities as follows.

A: Literal pairs

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

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

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

The first inequalities in B, together with those in A, imply $T[v, h_i] \leq 1$ for all i, from which we can then use the second inequalities to deduce that $T[v, h_i] = 1$ for all i. The vertex v must be at the center of a star

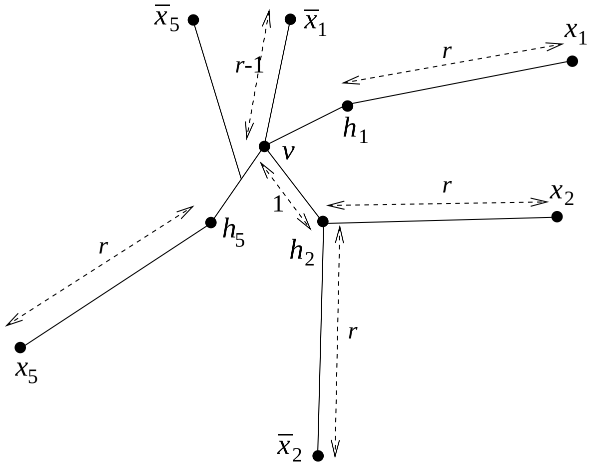

Figure 5.2: Portion of sample layout

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. The other path may do likewise or may begin towards v for a distance of up to 1 before going away from v. Such a path cannot follow any edge towards an h_j $(j \neq i)$ because of the inequality $T[h_i, \tilde{x}_j] \geq r$. An impression of a general feasible configuration is presented in Figure 5.2.

The essential feature of such configurations, which we shall take advantage of in our reduction, is that for each i, *at most one* of x_i and $\bar{x}_i$ is within distance $r - 1$ of v. At least one of them is at distance $r + 1$. The final inequalities will represent the satisfaction of clauses by literals. A satisfying literal will correspond to a vertex $\tilde{x}_i$ which is within $r - 1$ of v.

Now, we present the third set of inequalities that deal with the "clause" vertices c_j, c'_j, c''_j.

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).

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

$$T[c', y''] \le r+1,\ T[c', y] \le r+1,$$
$$T[c'', y] \le r+1,\ T[c'', y'] \le r+1,$$
$$T[c, c'] \ge 2,\ T[c', c''] \ge 2,\ T[c'', c] \ge 2.$$

We have argued already that if inequalities A and B are satisfied then v must be on the path between any two literals of distinct variables. 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 within $r-1$ of v then a configuration of the form illustrated in Figure 5.3 is feasible.

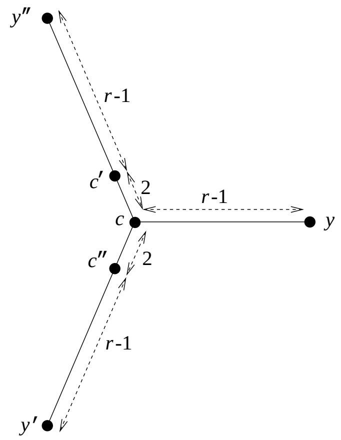

Figure 5.3: Layout of clause vertices

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 it is easy to see that there is a distance matrix D such that the approximation to it within $\Delta = 2$ has the same effect as the set of inequalities for T we have presented. For example,

$$D[v, \tilde{x}_i] = r-1, D[v, h_i] = 0, D[v, c] = 2$$

for all i and clause vertices c. The remaining entries of D can be filled similarly. ■

Theorem 5.2 *It is an $\mathcal{NP}$-hard problem, given a distance matrix D, to find an additive metric T such that*

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

Proof: It should be noted that the above proof already carries some non-approximability. The $\frac{9}{8}$ factor is achieved by some slight modifications. ■

6 Generalization to Other Norms

First, we show that Lemma 3.3 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 3.3 returns an a-restricted tree $T^{/a}$ such that

$$\forall i, j, |T^{/a}[i,j] - D[i,j]| \leq |\varepsilon[i,j]| + |\varepsilon[a,i]| + |\varepsilon[a,j]|. \tag{3}$$

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

$$\sum_{i=1}^{k} \frac{|x_i|^p}{k} \geq \left| \frac{\sum_{i=1}^{k} x_i}{k} \right|^p. \tag{4}$$

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 (3) and (4) 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 dissimilairty. In D. Kendall and P. Tautu, editors, *Mathematics in Archeological and Historial Science*, 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*, 1993. In press. See also STOC '93.

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