Research Report 330

On the Approximability of Numerical Taxonomy (Fitting Distances by Tree Metrics)

Richa Agarwala, Vineet Bafna, Martin Farach, Mike Paterson and Mikkel Thorup

RR330

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 a tree metric T such that $\| T - D \|_\infty \leq 3\varepsilon$. Second we show that it is NP– hard to find a tree metric T such that $\| T - D \|_\infty < 9/8\varepsilon$. This paper presents the first algorithm for this problem with a performance guarantee.

Department of Computer Science
University of Warwick
Coventry CV4 7AL
United Kingdom

July 1997

ON THE APPROXIMABILITY OF NUMERICAL TAXONOMY (FITTING DISTANCES BY TREE METRICS)

RICHA AGARWALA*, VINEET BAFNA†, MARTIN FARACH‡, MIKE PATERSON§, AND MIKKEL THORUP¶

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 a tree metric T such that $\| T - D \|_\infty \leq 3\varepsilon$. Second we show that it is $\mathcal{NP}$-hard to find a tree metric T such that $\| T - D \|_\infty < \frac{9}{8}\varepsilon$. This paper presents the first algorithm for this problem with a performance guarantee.

Key words. Approximation algorithm, tree metric, taxonomy.

AMS subject classifications. 62P10, 68Q25, 92B10, 92-08.

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

*DIMACS, Rutgers University, Piscataway, NJ 08855, USA. Supported by Special Year National Science Foundation grant BIR-9412594. Current address: National Human Genome Research Institute/National Institutes of Health, Bethesda, MD 20892, (richa@helix.nih.gov)

†(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.

§Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK. (msp@dcs.warwick.ac.uk) Supported in part by the ESPRIT LTR Project no. 20244 — ALCOM-IT.

¶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.

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* metrics. There may be no tree metric 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 $\| T - D \|_k$ ($\| T - D \|_k$ is formally defined in Definition 2.6).

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 metric T coinciding exactly with D, it is unique and constructible in linear, i.e., $O(|S|^2)$, time. Unfortunately there is typically no tree metric 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 additive 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 metric 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 an additive metric 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 with the same loss of a factor of 3.

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 L_k 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 ([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 on n objects 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*

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

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

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 any point a, D is quasi-additive if and only if $D + C^a$ is an 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 and let $\mathcal{X}$ be the set of all additive metrics. We define $\mathcal{A}(D)$ to be (one of) the additive metrics such that

$$\| D - \mathcal{A}(D) \|_\infty = \min_{A \in \mathcal{X}} \| D - A \|_\infty$$

For point a, we say a metric M is *a-restricted* if $\forall i, M[a, i] = D[a, i]$. Let $\mathcal{X}^a$ be the set of a-restricted additive metrics. We define $\mathcal{A}^a(D)$ to be (one of) the a-restricted additive metrics such that $\| D - \mathcal{A}^a(D) \|_\infty = \min_{A \in \mathcal{X}^a} \| D - A \|_\infty$. In other words, $\mathcal{A}^a(D)$ is an optimal a-restricted additive metric for D. We will sometimes refer to such a metric as *a-optimal.* Similarly, we define $\mathcal{U}(D)$ to be an optimal ultrametric for D. Note that the functions, $\mathcal{A}(), \mathcal{A}^a()$, and $\mathcal{U}()$, need not be uniquely-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 [6].

THEOREM 3.3. *Consider two $n \times n$ distance matrices $L, M : S^2 \to \Re_{>0}$ and a real value h such that $L[i,j] \leq M[i,j] \leq h$ for all i, j. There is an $O(n^2)$ algorithm to compute an ultrametric U, if it exists, such that for all i, j, $L[i,j] \leq U[i,j] \leq h$, and $\| M - U \|_\infty$ is minimized.*

Proof. Our proof uses the construction of Theorem 5 in [6]. First we show how, given a distance matrix $A : S^2 \to \Re_{>0}$, we can construct in time $O(n^2)$ an ultrametric U, such that $U \leq A$ (i.e., $\forall i, j : U[i,j] \leq A[i,j]$) and such that for any ultrametric $U' \leq A$, $U' \leq U$.

Let T be a minimum spanning tree over the graph defined by A. The ultrametric U is now defined as follows. Let $e = (i,j)$ be the maximum weight edge of T, and let T_1 and T_2 be the subtrees of T obtained by deleting (i,j). Then U has root at height $A[i,j]/2$ and the subtrees of the root are the ultrametric trees U_1 and U_2 recursively defined on T_1 and T_2. Clearly, $U \leq A$.

CLAIM: 3.3.1. *For any ultrametric U', if $U' \leq A$ then $U' \leq U$.*

PROOF: Let S_1 and S_2 be the partition of S defined by T_1 and T_2. By induction, for $k = 1, 2$, $U_k \geq U'|S_k^2$.

Let U_1' and U_2' be the two subtrees of U, and let S_1' and S_2' be the corresponding partitioning of S. Set $w = A[i,j]$ and $w' = \min_{(i,j) \in S_1' \times S_2'} A[i,j]$. Since w is the maximum weight in the minimum spanning tree T, $w' \leq w$. However, it is required that $U' \leq A$, so the height of the root of U' is $w'/2$, that is, the maximal distance in U' is w'. Thus for all $(i,j) \in S_1 \times S_2$, $U[i,j] = w \geq w' \geq U'[i,j]$. □

Consider an ultrametric U' as described in the Theorem 3.3, i.e. for all i, j, $L[i,j] \leq U'[i,j] \leq h$, and $\varepsilon = \| M - U' \|_\infty$ is minimized. Set

$$\varepsilon^+ = \max_{i,j \in S}, (M[i,j] - U'[i,j]) \leq \varepsilon$$

Suppose that we knew ε^+. Define A^{ε^+} such that $A^{\varepsilon^+}[i,j] = \min\{M[i,j] + \varepsilon^+, h\}$, and construct T^{ε^+} and $U^{\varepsilon^+} \leq A^{\varepsilon^+}$ as described above. Since $U' \leq A^{\varepsilon^+}$, $U^{\varepsilon^+}[i,j] \geq U'[i,j]$, so $L[i,j] \leq U^{\varepsilon^+}[i,j]$ and $\| M - U^{\varepsilon^+} \|_\infty \leq \| M - U' \|_\infty$.

Now observe that if T is a minimum spanning tree for M then T is also a minimum spanning tree for A^{ε^+}. Thus it follows that the topology of an optimal ultrametric U may be the same as the one we would construct from T and M. Given that T defines the right topology, we can construct the optimal ultrametric as follows.

Let $e = (i,j)$ be the maximum M-weight edge of T, and let T_1 and T_2 be the subtrees of T obtained by deleting (i,j). Let S_1 and S_2 be the partition of S defined by T_1 and T_2. Set

$$\mu = \frac{\max_{(i,j) \in S^2} M[i,j] + \min_{(i,j) \in S_1 \times S_2} M[i,j]}{2}$$

Then U has root at height $\min\{h, \mu\}/2$ and the subtrees of the root are the ultrametric trees U_1 and U_2 recursively defined on T_1 and T_2. □

3.1. The L_∞ Approximation. The *stem* of a leaf is the edge incident on it.

LEMMA 3.4. *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. We will move all i either towards or

away from a until each i is distance $D[a,i]$ from a. If $\mathcal{A}(D)[a,i] - D[a,i]$ is negative, we simply increase the length of its stem. Otherwise, i must be moved closer to a. Consider the (weighted) path from i to a. Let p be the point on this path which is distance $D[a,i]$ from a. We simply move i to this location. In either case, no point i is moved more than $|\varepsilon[a,i]|$, and so $|\mathcal{A}(D)[i,j] - T'^{a}[i,j]| \leq |\varepsilon[a,i]| + |\varepsilon[a,j]|$. 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 3.5. *For any point a, $\mathcal{A}^a(D)$ 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:

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

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

For any distance matrix M, define $\mathcal{U}^a(M)$ to be an a-restricted ultrametric which minimizes $\| M - \mathcal{U}^a(M) \|_\infty$. Note that for all i,j, $\mathcal{U}^a(M)[i,j] \leq 2m_a$. We can therefore apply Theorem 3.3, and so $\| M - \mathcal{U}^a(M) \|_\infty$ 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: 3.5.1. *T is an a-restricted additive metric.*

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

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

Now, since $\mathcal{U}^a(D^a)$ is an 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: 3.5.2. *$\mathcal{A}^a(D) + C^a$ is an a-restricted ultrametric.*

.PROOF: From Lemma 3.1, $\mathcal{A}^a(D) + C^a$ is an ultrametric. To show that Constraint (2) is satisfied, define $T' = \mathcal{A}^a(D) + C^a$ and 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 use the fact that $\mathcal{A}^a(D)$ is a metric, and therefore, for all $i,j \neq a$,

$$\begin{aligned} & \mathcal{A}^a(D)[a,j] \le \mathcal{A}^a(D)[i,j] + \mathcal{A}^a(D)[a,i] \\ \Rightarrow \quad & T'[a,j] - C^a[a,j] \le T'[i,j] - C^a[i,j] + T'[a,i] - C^a[a,i] \\ \Rightarrow \quad & T'[a,j] \le T'[j,i] + T'[a,i] - 2l_i \\ \Rightarrow \quad & 2m_a \le T'[j,i] + 2m_a - 2l_i \\ \Rightarrow \quad & T'[j,i] \ge 2l_i. \end{aligned}$$

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

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

Therefore, Constraint (1) is also satisfied and Claim 3.5.2 is proved. □

Finally,

$$\begin{aligned} \| T - D \|_\infty &\ge \| \mathcal{A}^a(D) - D \|_\infty \text{ (by Claim 3.5.1)} \\ &= \| (\mathcal{A}^a(D) + C^a) - (D + C^a) \|_\infty \\ &\ge \| \mathcal{U}^a(D + C^a) - (D + C^a) \|_\infty \text{ (by Claim 3.5.2)} \\ &= \| T - D \|_\infty \text{ (by construction).} \end{aligned}$$

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

Lemmas 3.4 and 3.5 imply:

THEOREM 3.6. *Given an $n \times n$ distance matrix D, we can find a tree metric T in $O(n^2)$ time such that*

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

4. Tightness of analysis. In this section we show that the constant in Lemma 3.4 is tight, and that for some distance matrices it is not improved by trying different values of c.

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

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

Proof. Consider the following distance matrix D for the points $q_0, \cdots, q_8$:

$$\begin{aligned} D[q_i, q_j] &= d - \varepsilon \quad \text{if } i = (j+1) \bmod 9 \text{ or } i = (j-1) \bmod 9 \\ &= 0 \quad \text{if } i = j \bmod 3 \\ &= d + \varepsilon \quad \text{otherwise.} \end{aligned}$$

Note that for each $c = q_i$, there exists $a_1 = q_{(i+1) mod 9}$, $a_2 = q_{(i-1) mod 9}$, $b_1 = q_{(i+4) mod 9}$, and $b_2 = q_{(i-4) mod 9}$ such that

$$\begin{aligned} D[c,a_1] &= D[a_2,c] = D[b_2,b_1] = d - \varepsilon, \\ D[b_1,c] &= D[c,b_2] = D[a_1,a_2] = d + \varepsilon, \text{ and} \\ D[a_1,b_1] &= D[a_2,b_2] = 0. \end{aligned}$$

If we take d to be much larger than ε, it is easy to see that any reasonable approximation by a tree metric T uses a tree with a central vertex m from which three edges lead to subtrees containing c, $\{a_1, b_1\}$ and $\{a_2, b_2\}$ respectively.

Hence,

$$T[b_1, c] - T[c, a_1] + T[a_1, a_2] - T[a_2, c] + T[c, b_2] - T[b_2, b_1] = 0,$$

whereas

$$D[b_1, c] - D[c, a_1] + D[a_1, a_2] - D[a_2, c] + D[c, b_2] - D[b_2, b_1] = 6\varepsilon.$$

Therefore any such approximation T satisfies $\| D - T \|_\infty \geq \varepsilon$.

For a c-restricted approximation T (where $T[u, c] = D[u, c]$ for all c), we find that

$$T[a_1, a_2] - D[a_1, a_2] - T[b_2, b_1] + D[b_2, b_1] = 6\varepsilon,$$

and so $\| D - T \|_\infty \geq 3\varepsilon$.

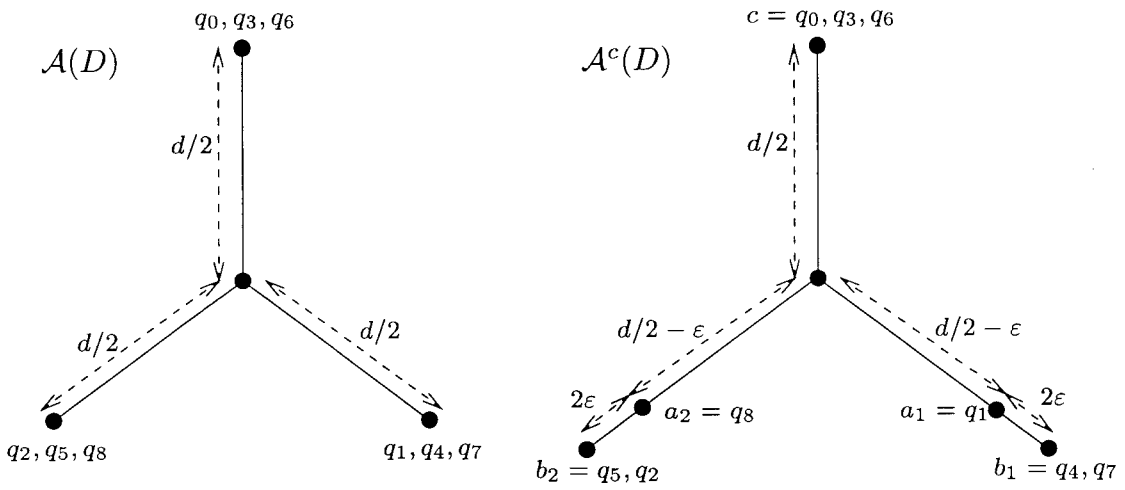

FIG. 1. *Trees approximating D.*

Figure 1 shows optimal solutions which establish that $\| D - \mathcal{A}(D) \|_\infty = \varepsilon$ and that $\| D - \mathcal{A}^c(D) \|_\infty = 3\varepsilon$. □

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. First, we show that a decision version of the Numerical Taxonomy Problem is $\mathcal{NP}$-complete.

The Numerical Taxonomy Problem

Input: A distance matrix $D : S^2 \to \Re_{\geq 0}$, 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. That the problem is in $\mathcal{NP}$ is immediate. 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 $x_i, \bar{x}_i, h_i$ for $1 \leq i \leq n$, and c_j, c'_j, c''_j for $1 \leq j \leq k$, and v.

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 $D[x_i, \bar{x}_i] = 2r + \Delta$. We classify the inequalities as follows.

A: *Literal pairs*

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

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*

$$(1) \quad T[v, x_i] \leq r + 1, \quad T[v, \bar{x}_i] \leq r + 1, \quad \text{for all } i,$$

$$(2) \quad T[h_i, h_j] \geq 2, \quad T[h_i, x_j] \geq r, \quad T[h_i, \bar{x}_j] \geq r, \quad \text{for all } i, j \ (i \neq j).$$

The inequalities B(1), together with those in A, imply $T[v, h_i] \leq 1$ for all i, and we can then use the first inequality of B(2) to deduce that $T[v, h_i] = 1$ for all i.

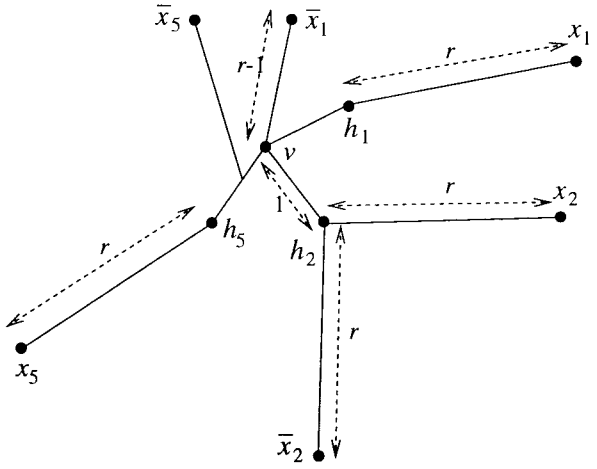

FIG. 2. *Portion of sample layout*

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 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 $\tilde{x}_i \in \{x_i, \bar{x}_i\}$ 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 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 j 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 last three 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 3 is feasible.

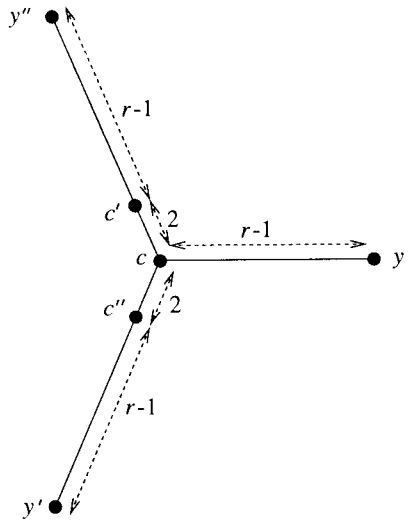

FIG. 3. *Layout of clause vertices*

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.

To complete the proof, we construct a distance matrix D such that (1) if for some tree metric T, $\| T - D \|_\infty \leq \Delta$, then T satisfies all the inequalities from A, B and C, and (2) for the tree layout T described above, corresponding to a satisfiable 3SAT expression, we have $\| T - D \|_\infty \leq \Delta$.

Concerning (1), for all vertices a, b, and all $z \in \Re_{\geq 0}$, if an inequality is of the form $T[a, b] \geq z$, let $D[a, b] = z + \Delta$. Correspondingly, if the inequality is of the form

$T[a,b] \leq z$, let $D[a,b] = z - \Delta$. Concerning (2), for $\tilde{x}_i \in \{x_i, \bar{x}_i\}$, $\tilde{x}_j \in \{x_j, \bar{x}_j\}$, $i \neq j$, in our intended configurations we have $2r - 2 \leq T[\tilde{x}_i, \tilde{x}_j] \leq 2r + 2$, with either extreme possible. Therefore we take $D[\tilde{x}_i, \tilde{x}_j] = 2r$. Since $\Delta = 2$, this covers both extremes. Similarly, for each clause C we take $D[c,y] = D[c',y'] = D[c'',y''] = r + 1$. Suitable values for the remaining entries of D are easy to find. This completes the proof of Theorem 5.1. □

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 some of the inequalities by a fixed amount δ and omitting others. The matrix D is the same as before.

A: *Literal pairs*

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

B: *Star-like tree*

$$T[v, x_i] \leq r + 1 + \delta, \quad T[v, \bar{x}_i] \leq r + 1 + \delta, \quad T[v, h_i] \leq 1 + \delta, \quad \text{for all } i,$$

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

C: *Clause satisfaction*

For each clause $C = (y, y', y'')$ where y, y', y'' are literals, we have three vertices c, c', c'' and the following inequalities:

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

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

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

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

Note that 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 consider an arbitrary tree T which satisfies inequalities A, B and C. Our aim will be to show that if S is not satisfiable then $\delta \geq 1/4$, and so $\| D - T \|_\infty \geq 2 + 1/4$.

For any three distinct tree vertices u, v, w, let $\texttt{meet}(u,v,w)$ denote the intersection point of the paths between them. We interpret x_i as false if and only if $T[h_i, \texttt{meet}(v, h_i, x_i)] \leq T[h_i, \texttt{meet}(v, h_i, \bar{x}_i)]$. Without loss of generality, we may *restrict our attention to a tree for which our interpretation sets all x_i to be false.*

For any variable x_i, let $\hat{h}_i$ denote $\texttt{meet}(h_i, x_i, \bar{x}_i)$, and $\hat{v}_i$ denote $\texttt{meet}(v, h_i, x_i)$. Note that x_i being false implies that $T[h_i, \hat{v}_i] \leq T[h_i, \hat{h}_i]$.

CLAIM: 5.2.1. *For all i, (1) $T[h_i, \hat{v}_i] \leq T[h_i, \hat{h}_i] \leq 3\delta/2$, and (2) $T[x_i, \hat{h}_i] - T[h_i, \hat{h}_i] \geq r - 2\delta$.*

PROOF: For (1), $2T[h_i, \hat{h}_i] = T[x_i, h_i] + T[\bar{x}_i, h_i] - T[x_i, \bar{x}_i] \leq 2(r+\delta) - (2r-\delta) = 3\delta$.

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

For each $j \neq i$, set $h_i^j = \texttt{meet}(h_j, h_i, x_i)$.

CLAIM: 5.2.2. *For all $\delta < \frac{2}{7}$ and for all $j \neq i$, $T[h_i, h_i^j] < T[h_i, \hat{v}_i]$.*

PROOF: Suppose $T[h_i, h_i^j] \geq T[h_i, \hat{v}_i]$. Then there are simple paths from h_i to $\hat{v}_i$ to h_j and from v to $\hat{v}_i$ to h_j. Therefore

$$\begin{aligned} 0 &= T[h_i, \hat{v}_i] + T[\hat{v}_i, h_j] - T[h_i, h_j] \\ &\leq T[h_i, \hat{v}_i] + T[v, h_j] - T[h_i, h_j] \\ &\leq 3\delta/2 + (1+\delta) - (2-\delta) \end{aligned}$$

Hence $\delta \geq \frac{2}{7}$. □

CLAIM: 5.2.3. *For all $\delta < \frac{2}{7}$ and for all $i \neq j$, $T[x_i, x_j] \geq 2r + 2 - 5\delta$.*

PROOF: By Claims 5.2.2 and 5.2.1(1), $T[h_i, h_i^j] \leq 3\delta/2$ and $T[h_j, h_j^i] \leq 3\delta/2$. Since $T[h_i, h_j] \geq 2 - \delta$ and $\delta \leq 1/2$, we may conclude that we have a simple path from h_i to h_i^j to h_j^i to h_j, and a simple path from x_i to $\hat{h}_i$ to h_i^j to h_j^i to $\hat{h}_j$ to x_j. Note, however, that $\hat{h}_i$ and h_i^j may coincide, and similarly for h_j^i and $\hat{h}_j$. In conclusion,

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

For the last inequality, we used Claim 5.2.1(2). □

Finally, we show that if S is not satisfiable then $\delta \geq 1/4$. If $\delta \geq 2/7$ then this is trivially true, so we may assume that the conclusions of Claims 5.2.2 and 5.2.3 apply.

Let vertices x, x', x'' in T correspond to the three false literals of a clause. Let $p = \texttt{meet}(x, x', x'')$. Without loss of generality, assume $T[x, p] \geq T[x', p] \geq T[x'', p]$. Let d be at the middle of the path from x to x''. By Claim 5.2.3, $T[x, d] = T[x'', d] \geq r + 1 - 5\delta/2$. Hence the bounds of $r+1+\delta$ on $T[x, c']$ and $T[x'', c']$ from the inequalities in C imply that $T[d, c'] \leq 7\delta/2$.

Now d is situated on the path from x to p, and $T[p, x'] \geq T[p, x'']$, implying $T[d, x'] \geq T[d, x''] \geq r + 1 - 5\delta/2$. Hence, as above, the bounds of $r + 1 + \delta$ on $T[x, c'']$ and $T[x'', c'']$ imply that $T[d, c''] \leq 7\delta/2$. Consequently $T[c', c''] \leq T[c', d] + T[d, c''] \leq 7\delta$. However, from C we also have the inequality $T[c', c''] \geq 2 - \delta$. Thus $7\delta \geq 2 - \delta$ and so $\delta \geq 1/4$.

Since T was arbitrary, we have shown that if S is not satisfiable then there is no tree T such that $\| D - T \|_\infty < 2 + 1/4$, i.e., $\| D - \mathcal{A}(D) \|_\infty \geq 2 + 1/4$. □

Theorem 5.2 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*

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

Proof. For any such T, if $\| D - T \|_\infty \geq 2 + 1/4$ then $\| D - \mathcal{A}(D) \|_\infty > 2$ and S is unsatisfiable, and if otherwise then $\| D - \mathcal{A}(D) \|_\infty \leq \| D - T \|_\infty < 2 + 1/4$ and S is satisfiable. □

6. Generalization to Other Norms. First, we show that Lemma 3.4 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 3.4 returns an a-restricted tree $T^{/a}$ such that

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

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. 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 $\| D - U \|_k$.

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 $T(n)$ on an $n \times n$ matrix, then there is an algorithm $F(D)$ which achieves a 3α-approximation for the Additive$_k$ problem and runs in $O(nT(n))$ 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 Mathematics in the Archaeological and Historical Sciences, F. Hodson, D. Kendall and P. Tautu, editors, Edinburgh University Press, Edinburgh, 1971, pp. 387–395.

[4] L. CAVALLI-SFORZA AND A. EDWARDS, *Phylogenetic analysis models and estimation procedures*, Amer. J. Human Genetics, 19 (1967), pp. 233–257.

[5] W.H.E. DAY, *Computational complexity of inferring phylogenies from dissimilarity matrices*, Bulletin of Mathematical Biology, Vol. 49, No. 4 (1987), pp. 461–467.

[6] M. FARACH, S. KANNAN, AND T. WARNOW, *A robust model for finding optimal evolutionary trees*, Algorithmica, 13 (1995), pp. 155–179.

[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 Molecular Systematics, D. M. Hillis and C. Moritz, editors, Sinauer Associates Inc., Sunderland, MA., 1990, pp. 411–501.

[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 (1977), pp. 199–213.