15-859(J): Metric Embeddings and Methods **Lecturer:** R. Ravi
Topic: Ultrametrics Continued and Closest Metric Problems **Date:** Oct. 14, 3002
Scribe: Daniel Golovin

13.1 Ultrametrics, Continued

Recall the definition of an Ultrametric:

Definition 13.1.1 *An Ultrametric is a metric (X, d) such that for all triangles in X, the maximum edge length in the triangle is non-unique.*

Via induction, this implies that the maximum edge length in any cycle in G, the graph generating (X, d), is non-unique. It also implies that the distance between two points x and y in an ultrametric (X, d) is the length of the longest edge in the xy path in any minimum

spanning tree (MST) of X under d. That is, $(\forall x, y)(d(x, y) = \max_{(u,v)\in P} d(u, v))$ where the sum is taken over edges in the i, j path P of some fixed MST of G. Note that any MST will do. Finally, we say an ultrametric (X, U) is *associated* with a graph G if it is generated by MST(G) in this way.

13.2 The Sub-Dominant Ultrametric

We now proceed to define the"closest" ultrametric smaller than a given distance matrix (not necessarily a metric). Note we can treat $M \in \mathbb{R}^{n\times n}$ as a graph $G = K_n$ with edge (a, b) weighted by $M(a, b)$.

Definition 13.2.1 *Given a distance matrix M on a set X, such that M does not necessarily obey the triangle inequality, define the Sub-Dominant Ultrametric $S_M = (X, U)$ of M as the ultrametric associated with (the graph form of) M.*

That is, if $T = \text{MST}(M)$ and $Path_T(i, j)$ is the unique ij path in T then

$$U(i, j) = \max_{(x,y)\in Path_T(i,j)} d(x, y)$$

Now we demonstrate that this ultrametric has the claimed properties.

Lemma 13.2.2 *The sub-dominant ultrametric $S_M = (X, U)$ is the unique ultrametric such that*

(1) $(\forall i, j)(U(i, j) \le M(i, j))$ *(which we abbreviate $U \le M$) and*

(2) For all ultrametrics $U' \le M$, $U' \le U$.

Proof: The proof has three parts, corresponding to the uniqueness of S_M, and claims (1) and (2) above.

(0) First observe that $S_M = (X, U)$ is uniquely generated by any MST of M. For suppose not. Then without loss of generality (WLOG) we have $(X, U), (X, U')$ generated by different MSTs T and T' such that for some pair $x, y \in X$, $U(x, y) < U'(x, y)$. Thus every edge in $\mathrm{Path}_T(x, y)$ has cost less than $U'(x, y)$. Removing from T' the most expensive edge in $\mathrm{Path}_{T'}(x, y)$, we get two connected components that can be connected by some edge in $\mathrm{Path}_T(x, y)$. Adding such an edge, we have reduced the total cost of T', contradicting the assumption that T' is an MST.

(1) Fix an MST of M, T, and vertices $x, y \in X$. If $e = (a, b)$ let $d(e) \equiv d(a, b)$, and similiarly define $U(e)$. Let $\mathrm{Path}_T(x, y) = \{e_1, \ldots, e_k\}$. Then $U(x, y) = \max_i M(e_i) \leq M(x, y)$.

(2) Consider another ultrametric $U' \leq M$. As before, let T be the MST generating U and let $\mathrm{Path}_T(x, y) = \{e_1, \ldots, e_k\}$. Then $U'(x, y) = \max_i U'(e_i) \leq \max_i M(e_i) = U(x, y)$. (Note that all maximums here are over edges of $\mathrm{Path}_T(x, y)$, including the term $\max_i U'(e_i)$.)

■

There are generalizations of this lemma for classes of metrics other than ultrametrics. These are the classes that obey the i^{th} power triangle inequality:

Definition 13.2.3 *The i^{th} order triangle inequality for distance matrix M over a set X is*

$$(\forall x, y, z \in X) \qquad (M(x, z))^i \leq (M(x, y))^i + (M(y, z))^i$$

Denote the set of metrics that obey the i^{th} power triangle inequality as $\mathcal{M}^i$

Note that the set of ultrametrics is equal to $\mathcal{M}^\infty$. We now proceed to show how the subdominant ultrametric fits into the larger context of closest metric problems.

13.3 Closest Metric Problems

In some instances, we may have a distance matrix M, not necessarily a metric, and we may wish to find the closest metric (X, d) to M according to some measure (possibly a metric on distance matrices), such that (X, d) is in a specified class $\mathcal{C}$ of metrics.

Definition 13.3.1 (The Closest Metric Problem) *Given distance matrix M, a set of metrics $\mathcal{C}$, and a distance function $D_M : \mathcal{C} \to \mathbb{R}_{\geq 0}$, find a metric $(X, d) \in \mathcal{C}$ minimizing D_M.*

As a concrete example, consider M given, $\mathcal{C}$ is the set of ultrametrics, and $D_M(U) = L_p(M, U) = [\sum_{ij} |M_{ij} - U(i, j)|^p]^{1/p}$ for some fixed value of p. We also define a point-fixed version of the closest metric problems, which will be easier to solve in some instances.

Definition 13.3.2 (The Point-Fixed Closest Metric Problem) *Given M, $\mathcal{C}$, and D_M as before, and a special vertex $a \in X$, find a metric $(X, d) \in \mathcal{C}$ minimizing D_M, such that $(\forall x \in X)$ $(d(a, x) = M_{ax})$. That is, find the closest metric $(X, d) \in \mathcal{C}$ in which all distances in (X, d) from some prespecified vertex 'a' agree with the input M.*

$\mathcal{C} \diagdown D_M$	L_1	L_2	L_∞	"Distortion"
Additive	NP-Hard$^+$	NP-Hard	NP-Hard*	NP-Hard
Ultrametric	NP-Hard$^+$	NP-Hard	P†	P
Path Metric	NP-Hard	NP-Hard	NP-Hard	NP-Hard$^+$

$\star$: There is a 3-approximation due to Agarwala et al.(1999). The vertex-fixed version in P.
$+$: Open problem, would be nice to approximate.
$\dagger$: Due to Farach, Kannan and Warnow (1995).

Figure 13.3.1: The Status of Closest Metric Problems

A table of results is given in Figure 13.3.1.

Theorem 13.3.3 (Farach, Kannan and Warnow (1995)) *The closest metric problem with the class $\mathcal{C}$ being the set of ultrametrics and the distance function $D_M(U) = L_\infty(M, U)$ is in P.*

Proof: This proof is due to Farach-Colton (2003). Consider $U \in \mathcal{C}$. For $M, N \in \mathbb{R}^{n\times n}$ let $[M, N] \equiv \{Q \in \mathbb{R}^{n\times n} \mid (\forall i, j)(M_{ij} \le Q_{ij} \le N_{ij})\}$ Let J be the matrix of all ones (i.e. for all i, j, $J_{ij} = 1$) of size n by n. Observe that

$$\Big(L_\infty(M, U) = \epsilon\Big) \Longleftrightarrow \Big(U \in [M - \epsilon J, M + \epsilon J]\Big) \Longleftrightarrow \Big(U \in [M - \epsilon J, S_{M+\epsilon J}]\Big)$$

However, for any ϵ, $S_{M+\epsilon J} = S_M + \epsilon J$, because S_M is formed from an MST of M. Thus

$$\Big(U \in [M - \epsilon J, S_{M+\epsilon J}]\Big) \Longleftrightarrow \Big(M - \epsilon J \le S_M + \epsilon J\Big) \Longleftrightarrow \Big(M - S_M \le 2\epsilon J\Big)$$

This occurs iff $\epsilon = \max_{i,j}\left(\frac{M(i,j)-S_M(i,j)}{2}\right)$ Therefore, the sub-dominant ultrametric S_M is the ultrametric closest to M with respect to the L_∞ measure. Since we can find S_M efficiently, this completes the proof. ■

13.4 Centroids and the Spherical Extension of a Tree

Definition 13.4.1 (X, d) *is a centroid metric if there exist lengths* $l_1, l_2, \ldots, l_n \ge 0$ *such that* $(\forall i, j)(d(i, j) = l_i + l_j)$*. Call* (X, d) *a centroid quasimetric if* $(\forall i, j)(d(i, j) = l_i + l_j)$ *for some (potentially negative) lengths* $l_1, l_2, \ldots, l_n$*.*

Proposition 13.4.2 *Given a tree (equivalently, additive) metric* (X, d) *and a centroid metric* (X, d')*,* $(X, d + d')$ *is a tree metric.*

Corollary 13.4.3 *Given a tree metric* (X, d) *and a centroid quasimetric* (X, d')*, if* $(X, d+d')$ *obeys non-negativity of distances and the triangle inequality, then it is a tree metric.*

Proof: First we give a proof though use of the 4 point property (4PC), then a constructive one by picture. Recall a metric has the 4PC iff the max weight matching on any four points is non-unique.

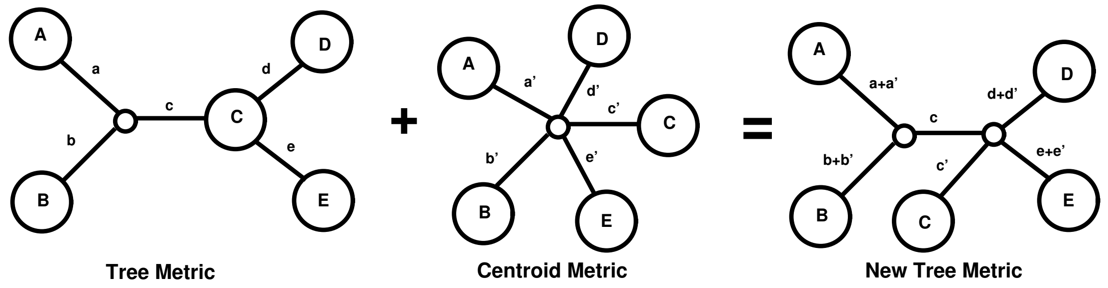

Figure 13.4.2: *An illustration of proposition 13.4.2.*

Recall also that additive/tree metrics obey the 4PC, as do centroid metrics. More strongly, we can say that all three matchings in any 4 point subset $S \subseteq X$ of a centroid metric have the same weight, namely $\sum_{i \in S} l_i$. Now fix any four points $x, y, w, z \in X$. Suppose WLOG that in (X, d) matchings $M_1 = \{(x, y), (w, z)\}$ and $M_2 = \{(x, z), (w, y)\}$ have the same (maximum) weight. Call M_3 the remaining matching. All three matchings will gain the same weight when we add d' to the distances. Thus, M_1 and M_2 will still be equal and maximal under $(X, d + d')$, and thus the 4PC holds for these points. Since we choose the points arbitrarily, it holds for all subsets, and thus $(X, d + d')$ has the 4PC and is thus a tree metric. Note the same argument holds in the case of centroid quasimetrics, except that we cannot gaurantee that $(X, d + d')$ obeys non-negativity of distances and the triangle inequality in general. ■

For the proof by picture, observe that a leaf node v in (X, d), we can add l_v to the length of the edge adjacent to v. For an internal node u, we can replace it by a new node w (i.e. a Steiner node) and add an edge (w, u) of length l_u. The resultant tree represents $(X, d + d')$.

13.4.1 Spherical Extension of a Tree

Definition 13.4.4 *Let (X, d) be a tree metric. The x-fixed centroid of (X, d), for $x \in X$ is denoted C^x and defined by $l_i = m_x - d(x, i)$ where $m_x = \max_i d(x, i)$. Thus, $C^x(i, j) = l_i + l_j = 2m_x - d(x, i) - d(x, j)$.*

Adding C^x to (X, d) is called "spherical extension of the tree."

Lemma 13.4.5 (The Interchange Lemma) *(X, D) is additive iff $(X, D) + C^x$ is ultrametric for all $x \in X$.*

Proof: We sketch a proof of only the forward direction. The general idea is to use the additive tree representation of (X, D), and add a few nodes and stretch existing edges to convert it into the tree represetation of an ultrametric. Recall the tree represetation of an ultrametric has the properties that (1) its leaves are exactly the set X, and (2) for every internal node v and every $x, y \in X$ that are descendents of v the length of the xv path equals the length of the yv path. Recall that these properties are sufficient to establish ultrametricity. Now fix $x \in X$, and T, a tree representation of (X, D). For leaves v of T, we add length l_v (obtained from C^x). For internal

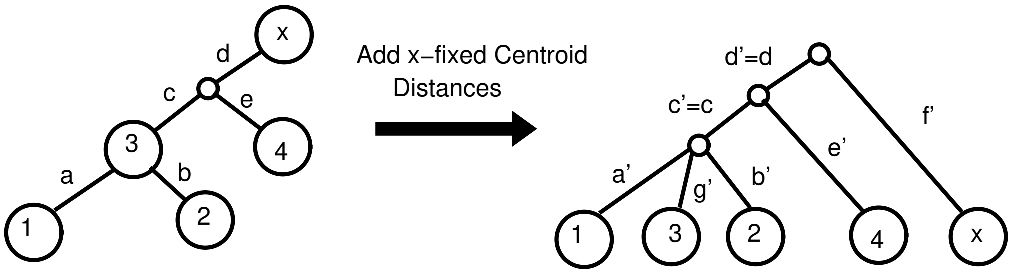

Figure 13.4.3: *The spherical extension of a tree metric to an ultrametric.*

nodes v and the root, we add a new node w where v was and add an edge (w, v) of length l_v.

Inspecting Figure 13.4.1, we see these operations obtain a tree representing an ultrametric. For example, consider $e' = e + m_x - D(x, 4) = e + m_x - (d + e) = m_x - d$. Thus the distance of vertex 4 to the root is equal to $e' + d = m_x$ in the new tree. ■

References

Agarwala, R., Bafna, V., Farach, M., Paterson, M., and Thorup, M. (1999). On the approximability of numerical taxonomy (fitting distances by tree metrics). *SIAM J. Comput.*, 28(3):1073–1085 (electronic).

Farach, M., Kannan, S., and Warnow, T. (1995). A robust model for finding optimal evolutionary trees. *Algorithmica*, 13(1-2):155–179.

Farach-Colton, M. (2003). (Nearly) Optimal Embeddings into Additive and Ultrametric Trees. *DIMACS workshop on discrete metric embeddings, Aug. 2003 (Princeton)*

Håstad, J., Ivansson, L., and Lagergren, J. (1998). Fitting points on the real line and its application to RH mapping. In *Algorithms—ESA '98 (Venice)*, volume 1461 of *Lecture Notes in Comput. Sci.*, pages 465–476. Springer, Berlin.