15-859(J): Metric Embeddings and Methods **Lecturer:** R. Ravi
Topic: Additive (Tree) Metrics **Date:** October 9, 2003
Scribe: Amitabh Sinha

12.1 Tree metrics

Definition 12.1.1 *A metric (X, d) is a* **tree** *metric if there is a tree T with non-negative edge lengths whose nodes contain X such that for every $x, y \in X$, we have $d(x, y) = d_T(x, y)$.*

Here $d_T(x, y)$ is the distance in the tree between x and y, defined as usual to be the sum of the lengths of the edges on the unique (x, y) path in the tree T. Tree metrics are also called **additive** metrics.

The tree T representing a metric may have points not in X. For example, the metric induced by K_n with all edge lengths 1 can be represented by a tree metric as follows: Construct a star with n leaves, one for each vertex in K_n, and assign length $1/2$ to each edge in the star. The center of the star is an extra node (which we will sometimes call Steiner nodes).

Since Steiner nodes of degree 2 can be removed without any loss, the number of Steiner nodes in a tree representation of (X, d) is no more than $|X| - 2$. Therefore, the metric (X, d) has a compact representation in the form of a tree with only $2|X| - 3$ lengths, as opposed to $\binom{n}{2} \approx n^2/2$ lengths needed to represent an arbitrary n-point metric.

12.1.1 Characterization of tree metrics

Definition 12.1.2 *A metric (X, d) is said to satisfy the* **four point condition (4PC)** *if for any four points, the maximum weight matching is not unique.*

The weight of a matching is the sum of distances between matched pairs of vertices. Given four points, there are exactly three distinct matchings possible on them. Therefore the 4PC says that at least two of them have the same weight which is no less than the weight of the third matching. In other words, if the four points are w, x, y, z, then there is a relabelling $\pi : \{w, x, y, z\} \to \{w, x, y, z\}$ such that $d(\pi(x), \pi(y)) + d(\pi(w), \pi(z)) \leq d(\pi(x), \pi(z)) + d(\pi(y), \pi(w)) = d(\pi(x), \pi(w)) + d(\pi(y), \pi(z))$.

Theorem 12.1.3 *A metric (X, d) is a tree metric if and only if it satisfies the four point condition.*

Proof: ($\Rightarrow$) Let T be a tree representing the tree metric (X, d), and let w, x, y, z be any four points in X. Consider the sub-tree induced by the paths $(w, x), (w, y)$ and (w, z) in T. Without loss of generality, it looks like the tree shown in Figure 12.1.1. Since all edges in the tree have non-negative lengths, we have $l \geq 0$. Therefore the matchings $\{(w, x), (y, z)\}$ and $\{(w, y), (x, z)\}$ have cost $l_1 + l_2 + l_3 + l_4 + 2l$, which is at least $l_1 + l_2 + l_3 + l_4$, which is the cost of the third matching, $\{(x, y), (w, z)\}$.

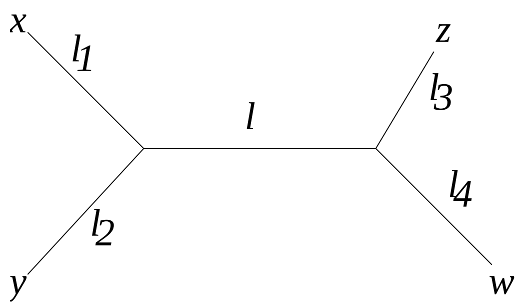

Figure 12.1.1: The four-point condition

($\Leftarrow$) The proof is by induction. Any metric with 3 or fewer points can trivially be represented by a tree metric. So we consider as the base case a metric with 4 points which satisfies 4PC. Without loss of generality, let the maximum weight matchings be $\{(w, x), (y, z)\}$ and $\{(w, y), (x, z)\}$. Then the tree shown in Figure 12.1.1 is the unique tree representing (X, d), with distances defined as follows: $l = (d_{xz} + d_{wy} - d_{xy} - dwz)/2$. Similarly, we have $l_1 = (d_{xz} + d_{xy} - dyz)/2$, and so on for l_2, l_3 and l_4. The fact that l_1, l_2, l_3 and l_4 are non-negative follows straight from the triangle inequality, without using 4PC. However, $l \geq 0$ follows from 4PC. Thus if 4PC holds for $\{w, x, y, z\}$, then these distances are well-defined and (X, d) can be represented as a tree metric.

A rough sketch of the induction is given below, with the details left out as an exercise for the reader. Suppose we have embedded into a tree $n - 1$ points of the n-point metric (X, d) which satisfied 4PC. Let w be the remaining point, and consider any three other points x, y, z. Let p be the unique degree-3 node in the subtree induced by x, y, z. Using the same argument as the base case, we can figure out which of the subtrees of p the node w belongs to. Once we know that, we again use the 4PC by selecting three other points in the subtree that w has been assigned to. Iterating this process eventually leads to fixing w exactly and gives us a tree representing (X, d). ∎

Corollary 12.1.4 *If (X, d) is additive, then there is a unique tree representing it.*

Proof: This follows from the proof of the sufficiency condition in Theorem 12.1.3. If $|X| \leq 3$, then clearly there is a unique tree representing (X, d). The rest of the proof proceeds by induction along the same lines as the proof in Theorem 12.1.3, which is a constructive proof with no degree of freedom at any point. ∎

Theorem 12.1.3 also gives an algorithm for construcing a tree representation of an additive metric (X, d). This algorithm has running time $O(n^2)$, since there are n nodes and we take linear time for each node. Consequently, it also gives an $O(n^2)$ time for checking whether or not (X, d) is additive, which is an improvement over the $O(n^4)$ time that checking 4PC for all four-tuples of points requires. In fact, additive metrics with constant degree can be reconstructed in $O(n \log n)$ time.

12.1.2 Centroid metrics

A very special class of tree metrics are *centroid* metrics, defined below. They will play an important role in the next class.

Definition 12.1.5 *A metric (X,d) is a* **centroid** *metric if the tree representing it is a star.*

12.2 Ultrametrics

Definition 12.2.1 *A metric (X,d) is said to satisfy the* **three point condition (3PC)** *if for any three points in X, the maximum distance between two of them is not unique. Metrics which satisfy 3PC are called* **ultrametrics***.*

The 3PC can be restated as: For every $x,y,z \in X$, we have

$$d_{xy} \leq \max\{d_{xz}, d_{yz}\}.$$

The following observation can be proved by repeated applications of 3PC.

Fact 12.2.2 *If (X,d) is an ultrametric, then for every path $x = t_0, t_1, \ldots, t_k = y$, we have*

$$d_{xy} \leq \max_{i=0}^{k-1}\{d_{t_i,t_{i+1}}\}.$$

Consider an ultrametric (X,d), and let T be its minimum spanning tree. Let x and y be a pair of points in X, and $x = t_0, t_1, \ldots, t_k = y$ be the path in T between x and y. By definition of MSTs, $d_{xy} \geq d_{t_i,t_{i+1}}$ for all i. However, since (X,d) is an ultrametric, we also have $d_{xy} \leq \max_i\{d_{t_i,t_{i+1}}\}$. Hence this must hold with equality, that is, $d_{xy} = \max_i\{d_{t_i,t_{i+1}}\}$ for every $x,y \in X$. This tree (the MST) is called the **generating** tree for the ultrametric (X,d).

This also gives us a fast algorithm for checking whether a given metric is an ultrametric or not. We construct the MST, and check if the distance between every pair of points in the original metric is equal to the length of the longest edge in the MST path between that pair of points. The following fact can be verified by an elementary case-analysis.

Fact 12.2.3 *If the metric (X,d) satisfies 3PC, then it also satisfies 4PC.*

Therefore, ultrametrics are tree metrics. In fact, the tree metrics satisfy a stronger condition.

Theorem 12.2.4 *If (X,d) is an ultrametric, then there is a tree metric T representing it rooted at a unique node r such that $d_T(r,x)$ is the same for every $x \in X$. This distance is known as the* **height** *of the tree.*

Proof: The proof is by induction. It is also a contructive proof. If $|X| = 2$, then creating a new node r and adding edges from r to each of the two points in X with length half the distance between the two points in X suffices. If $|X| = 1$ there is nothing to do. This completes the base case.

To construct the tree for (X, d) with $|X| > 2$, start with the MST M generating the ultrametric (X, d). Let $e = (x, y)$ be a longest edge in M. Delete e, and construct representing trees T_x and T_y for the two parts of M created by deleting e. Let T_x and T_y be rooted at r_x and r_y respectively, and let d_x and d_y represent the heights of T_x and T_y respectively.

Since (x, y) was the longest edge in the MST, $d_{xy} \geq 2d_x$ and $d_{xy} \geq 2d_y$. Create a new root node r, and connect it to r_x with an edge of length $\frac{d_{xy}}{2} - d_x \geq 0$. Connect r to r_y with an edge of length $\frac{d_{xy}}{2} - d_y \geq 0$. This preserves the distance between x and y in the tree, and also ensures that the height of the tree rooted at r is $d_{xy}/2$. ■

12.2.1 Applications

Perhaps the most striking and illustrative application of ultrametrics is in evolutionary biology. According to the theory of evolution, most life forms originated from a single life form through a series of genetic mutations. If we assume that the rate of genetic mutations is constant over time, then this process can be represented by an ultrametric tree with rooted at the originating life form and with all species alive today at the leaves.

If this hypothesis is true, then the "distance" between the genetic sequences of species forms an ultrametric. The distance can be measured in various ways, for example the Hamming distance between the DNA codes. The question of fitting this metric into an ultrametric now becomes very important, since such a fitting would reveal the evolutionary tree. Even an approximate fitting would give us useful information.

Formally, by fitting a metric M to a target metric family $\mathcal{T}$ we refer to the problem of finding $T \in \mathcal{T}$ which minimizes $|T - M|$, where $|.|$ is any matrix norm of our choice. In the next few classes, we will work towards a 3-approximation for the l_∞ matrix norm for fitting an arbitrary metric to an ultrametric. The application given above requires fitting the genetic sequence metric to an ultrametric.