ON EFFICIENT LOW DISTORTION ULTRAMETRIC EMBEDDING

Guillaume Lagarde

ICML'20

Joint work with

Vincent Cohen-Addad
CNRS, Google

Karthik C. S.
Tel-Aviv University

ULTRAMETRICS

✱ (X,d) is a metric space

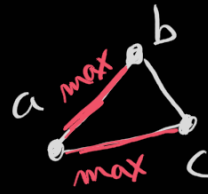
$$d(a,b) \leq d(a,c)+d(b,c)$$

✱ (X,Δ) is an ultrametric space

$$\Delta(a,b) \leq \max\{\Delta(a,c),\Delta(b,c)\}$$

✱ Interesting property

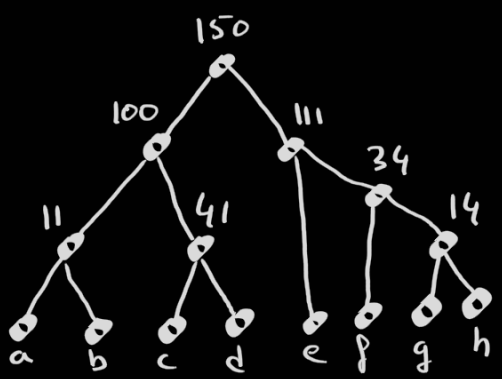
$\forall a,b,c \in X$



is isosceles "on the max"

+ extension to any cycle

ULTRAMETRIC = Tree



w : nodes $\longrightarrow \mathbb{R}^+$
non-increasing from root

$$\Delta(x,y) = w(LCA(x,y))$$

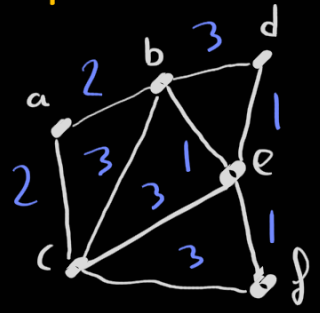
A few examples

Topology : discrete metric

Number Theory : p-adic numbers

$$d(x,y) = p^{-v_p(x-y)}$$

Graph Theory : minmax paths



$$\Delta(x,y) = \min_{p:x\rightsquigarrow y} \max_{e\in p} w(e)$$

$\Delta(b,c) = 2$

GOAL: embedding to ultrametric

Input

(X,d) a metric space

Output

Δ ultrametric s.t

$$d(x,y) \leq \Delta(x,y) \leq \rho_{OPT} \cdot d(x,y)$$

ρ_{OPT} ← minimal

called "max distortion"

$$\min_{\Delta} \left\| \frac{\Delta}{d} \right\|_\infty$$

with $\Delta \geq d$

$$\min_{\Delta} \left\| \frac{\Delta}{d} \right\|_p$$

p	Complexity	Approx
1	NP-hard APX-hard	$\log n$
2	NP-hard	$\sqrt{\log n \log\log n}$
p	? ? ?	$(\log n \log\log n)^{1/p}$
∞	$\Omega(n^2)$	

Results

Classic linkage algorithms

☹ time $O(n^2)$

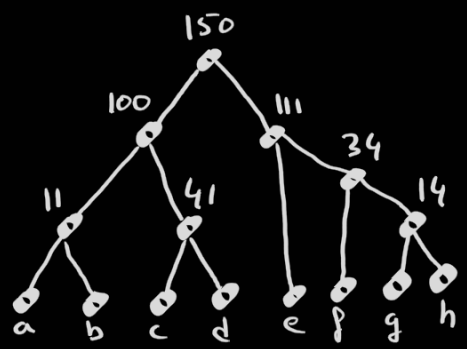
☹ space $O(n^2)$

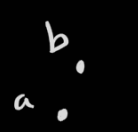
Theorem (Farach-Kannan-Warnow)

1. Optimal embedding in $O(n^2)$
2. Lower bound of $\Omega(n^2)$

Motivation: Unsupervised Learning

ultrametrics → hierarchical clustering





Algorithmica (1995) 13: 155–179

Algorithmica
© 1995 Springer-Verlag New York Inc.

A Robust Model for Finding Optimal Evolutionary Trees

M. Farach,[1] S. Kannan,[2] and T. Warnow[3]

Abstract. Constructing evolutionary trees for species sets is a fundamental problem in computational biology. One of the standard models assumes the ability to compute distances between every pair of species, and seeks to find an edge-weighted tree T in which the distance d_{ij}^T in the tree between the leaves of T corresponding to the species i and j exactly equals the observed distance, d_{ij}. When such a tree exists, this is expressed in the biological literature by saying that the distance function or matrix is *additive*, and trees can be constructed from additive distance matrices in $O(n^2)$ time. Real distance data is hardly ever additive, and we therefore need ways of modeling the problem of finding the best-fit tree as an optimization problem.

In this paper we present several natural and realistic ways of modeling the inaccuracies in the distance data. In one model we assume that we have upper and lower bounds for the distances between pairs of species and try to find an additive distance matrix between these bounds. In a second model we are given a partial matrix and asked to find if we can fill in the unspecified entries in order to make the entire matrix additive. For both of these models we also consider a more restrictive problem of finding a matrix that fits a tree which is not only additive but also *ultrametric*. Ultrametric matrices correspond to trees which can be rooted so that the distance from the root to any leaf is the same. Ultrametric matrices are desirable in biology since the edge weights then indicate evolutionary time. We give polynomial-time algorithms for some of the problems while showing others to be NP-complete. We also consider various ways of "fitting" a given distance matrix (or a pair of upper- and lower-bound matrices) to a tree in order to minimize various criteria of error in the fit. For most criteria this optimization problem turns out to be NP-hard, while we do get polynomial-time algorithms for some.

Theorem (Cohen-Addad, Karthik, L)

- For any $\gamma \geq 1$, $5 \cdot \gamma$-approx in time $O(n^{1+\frac{1}{\gamma^2}})$ for Euclidean metric
  - ↳ performs well on experiments!
- SETH ⇒ no $3/2$-approximate embedding in $n^{1.99}$ time from ℓ_∞-metric

Farach-Kannan-Warnow Algorithm

Input (X, d)

Output optimal ultrametric

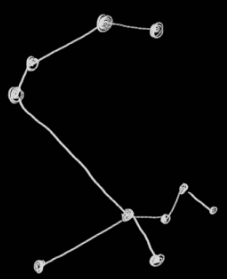
1. Compute a minimum spanning tree T
2. Compute cut weights of edges in T

$e = (a, b)$

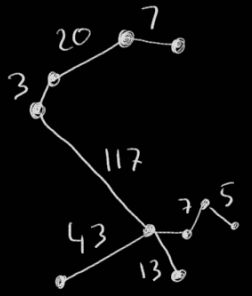
L_e = points accessibles from a by edges $< d(a,b)$

R_e = same with b

$$cw(e) = \max_{\substack{x \in L \\ y \in R}} d(x,y)$$



3. Compute a cartesian tree



Our Approx algorithm

1. Compute a γ-minimum spanning tree T
2. Compute β-cut weights of edges in T
3. Compute a cartesian tree

Claim this outputs a $\gamma \cdot \beta$-approx

Questions How to do 1. & 2. efficiently?

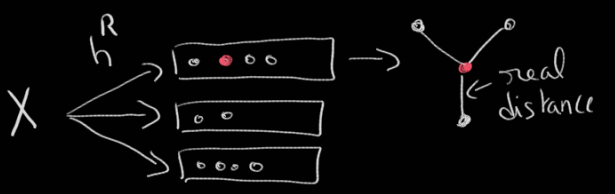
STEP 1 γ-MST time $nd + n^{1+1/\gamma^2}$

- $\forall \gamma \geq 1$, γ-spanner construction in $O(nd + n^{1+1/\gamma^2})$
  [HIS'13] Har-Peled Indyk Sidiropoulos

Locality sensitive hashing

$h^R : X \longrightarrow S$ "bucket space"

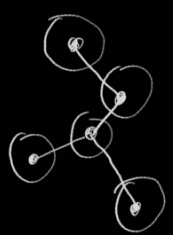
$h^R(p) = h^R(q)$ iff $d(p,q) < R$ (with high proba)



STEP 2. 5-approx for the cut weights | time $nd + n \log n$

Idea: tweak a union-find data structure

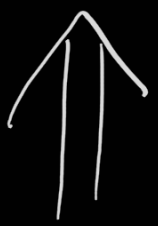
- create set
- merge sets
- query max distance between two sets



Hardness from SETH

Th [CKL] SETH $\Rightarrow$ no subquadratic algorithm can distinguish

$\mathbb{R}^{\log n}, \|\ \|_\infty$ | YES : $\exists$ isometric embedding
NO : Distortion $\geq 3/2$



Th [DKL'19] (bichromatic closest pair for $\|\ \|_\infty$)

SETH $\Rightarrow$ no subquadratic algorithm can distinguish, given sets $A, B \subseteq \mathbb{R}^{\log n}$

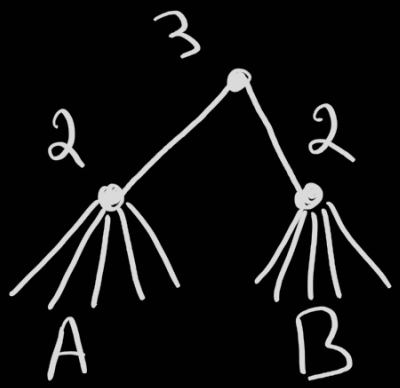
YES $\exists a, b$ s.t $\|a-b\|_\infty = 1$

NO $\forall a, b$, $\|a-b\|_\infty = 3$

Input $A, B \subseteq \mathbb{R}^{\log n}$

Promise $\forall\, a, a' \in A, \quad \|a - a'\|_\infty = 2$
$\forall\, b, b' \in B, \quad \|b - b'\|_\infty = 2$

YES case $\|A - B\|_\infty = 3$



$\longrightarrow$ isometric embedding

NO case $\exists\, a, a', b \quad s.t \quad \|a - b\| = 1$
$\|a' - b\| = 3$

$$3 = \|a' - b\|_\infty \leq \Delta(a', b)$$
$$\leq \max\{\Delta(a', a), \Delta(a, b)\}$$
$$\leq \max\{\rho \cdot \|a' - a\|_\infty, \rho \cdot \|a - b\|_\infty\}$$
$$\leq 2\rho$$

The last slide !

ADS → tool (with Rémi de Verclos)

github.com/guillaume-lagarde/fast-ultrametrics

Improved approximation factor ?

Cut-weights → enclosing ball ?

$5 \cdot \gamma \rightarrow \sqrt{3} \cdot \gamma$??

Euclidean Inapproxamibility under SETH ?

Improve the MST ?

(the true last slide)

Thank you!

FKW ALGO

def x, y ρ-separable
if $d(x,y) \leq \rho \cdot d(e_{x,y})$
↑ max edge on $x \rightsquigarrow y$

1. x, y not ρ-sep
$\Rightarrow \rho_{OPT} > \rho$ (no ρ-UM)

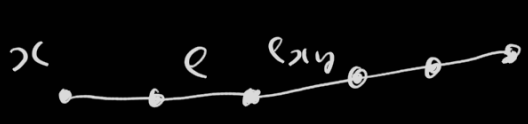
▷ 

$d(x,y) > \rho \cdot d(e_{x,y})$
$\geq \rho \cdot d(e)$ $\forall e \in x \rightsquigarrow y$

So unique max on the cycle
Contradiction

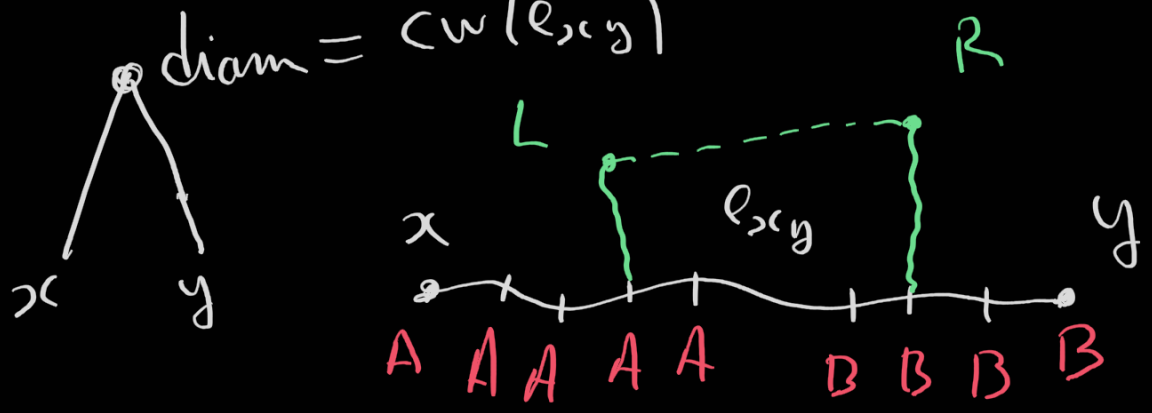
2. The construction give a ρ_{OPT}-UM for ρ_{OPT}.

$\Delta(x,y) \geq d(x,y)$
$=$
$cw(e)$



$cw(e) \geq cw(e_{x,y}) \geq d(x,y)$

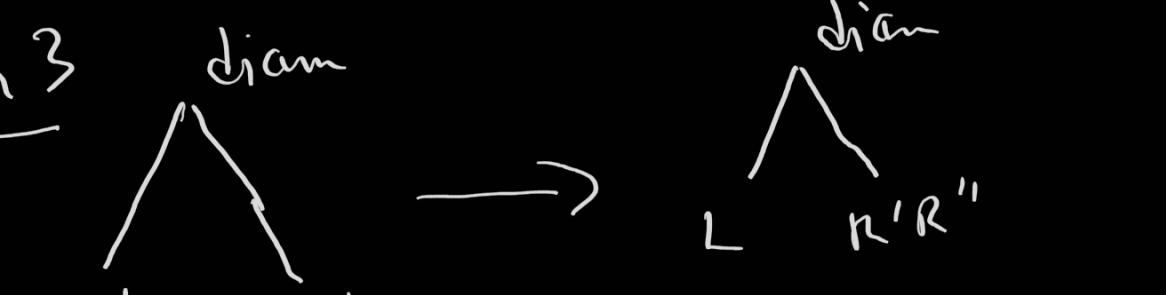
$\Delta(x,y) \leq \rho_{OPT} \cdot d(x,y)$



Claim 1 $\rho_{OPT} \geq \min_{A \cup B} \max_{\substack{a \in A \\ b \in B}} \frac{diam}{d(a,b)} = \frac{\Delta(a,b)}{d(a,b)}$

Claim 2 $\geq \frac{diam}{d(e_{xy})}$

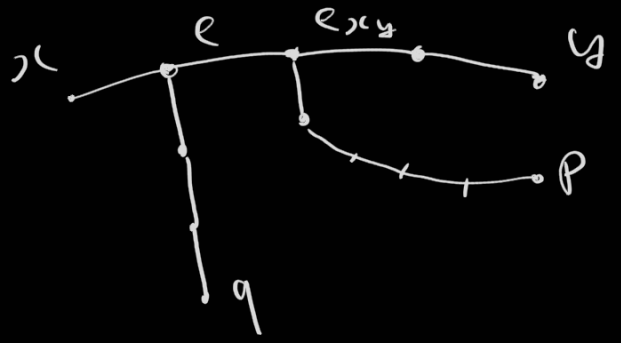
Claim 3



$dist(R', R'') \searrow$

$dist(L, R') \nearrow$

$\leq \frac{diam}{d(e_{xy})}$ (MST)



$$d(p, q) = cw(e)$$
$$\leq \rho_{OPT} \cdot \underbrace{d(e)}_{\leq d(x,y)}$$

because MST