Fitting Ultrametrics under any ℓ_p Objective

Moses Charikar* Ruiquan Gao*

November 3, 2026

Abstract

We study the ℓ_p-ultrametric fitting problem, where we are given pairwise distances $X : \binom{[n]}{2} \to \mathbb{R}_{>0}$ and the goal is to find an ultrametric $U : \binom{[n]}{2} \to \mathbb{R}_{>0}$ that minimizes $\|U - X\|_p$. We present a constant-factor approximation for any $p > 1$, improving upon the best-known $O((\log n \log\log n)^{1/p})$-approximation [Ailon and Charikar, SICOMP 2011]. This immediately implies a constant-factor approximation for the ℓ_p-tree fitting problem for any $p > 1$.

The algorithm is remarkably simple and reminiscent of classical average-linkage hierarchical clustering: we round the optimal solution of the standard LP relaxation by repeatedly merging clusters whose average inter-cluster LP distance is small, in a bottom-up manner. As a direct consequence of the analysis, we show the algorithm gives a 12.091-approximation for $p = 1$, improving upon the best-known 25.7846-approximation [An, Kao, Lee and Lee, FOCS 2025].

AI disclosure: These results were mostly found by GPT-6 Astra. The authors mainly integrated the results, verified and rewrote the proof, and improved the exposition.

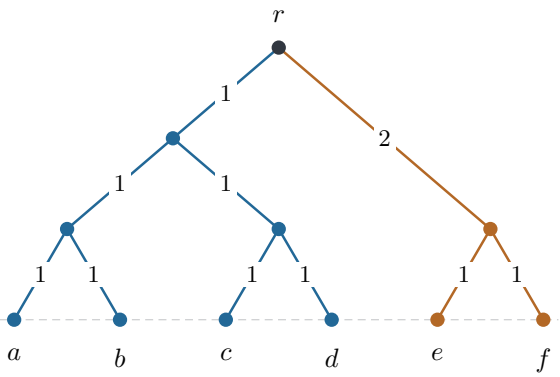
*Stanford University. {moses,ruiquan}@cs.stanford.edu

1 Introduction

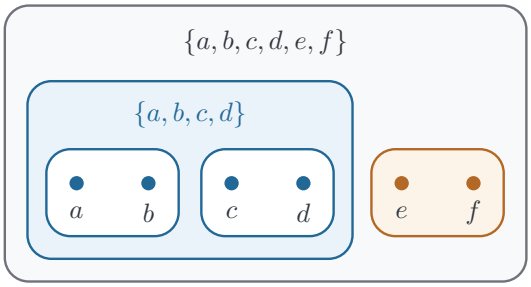
Ultrametrics provide a natural representation of hierarchical relationships among objects. An ultrametric U is defined as a metric satisfying a stronger ultrametric triangle inequality (implying that at least two edges share the maximum distance in any triangle):

$$U(i,j) \leq \max\{U(i,k), U(k,j)\} \ .$$

Equivalently, an ultrametric can be represented by the shortest path distances between leaves in a rooted tree in which the distances between the root and the leaves are equal; such a tree structure corresponds to a hierarchical clustering of the leaves.



(a) Tree strucutre of an ultrametric



(b) Corresponding hierarchical clustering

Because of this structure, ultrametrics have found applications in taxonomyony [Har67; Joh67; JJS67; JS71] for the classification of species, dating back to the 1960s. Since then, ultrametrics have been widely studied and used in many fields, including biology (e.g., [SS62; CE67]), physics (see the survey by [RTV86]), finance (e.g., [Man99]), and computer science.

In this paper, we study approximation algorithms for fitting an ultrametric to given measurements of distances between pairs of species, a problem which was first introduced by Cavalli-Sforza and Edwards [CE67] who proposed the least-squares formulation for the problem. Later on, Farris [Far72] proposed the problem of minimizing the ℓ_1 norm of the error, and these problems were shown to be NP-hard [KM86; Day87]. In the past three decades, this problem has attracted a lot attention in the TCS community, where minimizing the ℓ_p norm of the error (for various values of p) has been extensively studied. The current best-known results include a polynomial-time algorithm for the ℓ_∞ norm [FKW95], an 25.7846-approximation for the ℓ_1-norm [AKLL25] and an $O((\log n \log\log n)^{1/p})$-approximation for ℓ_p norms, $2 \leq p < \infty$ [AC11].

More specifically, in this paper, we study the ℓ_p-Ultrametric Fitting for any $1 \leq p < \infty$, formulated as follows: given n species and a matrix $X : \binom{[n]}{2} \to \mathbb{R}_{>0}$ representing pairwise dissimilarity estimates, our goal is to find an ultrametric $U : \binom{[n]}{2} \to \mathbb{R}_{>0}$ that minimizes the ℓ_p-distance to the input dissimilarity data X, i.e.,

$$\left(\sum_{i<j} |U(i,j) - X(i,j)|^p \right)^{1/p} .$$

Our main result is an $O(p)$-approximation for the ℓ_p-Ultrametric Fitting problem for any $p \geq 1$, which gives the first constant-factor approximation for any fixed $p > 1$.

Theorem 1. *For any $p > 1$, there exists an $O(p)$-approximation for the ℓ_p-Ultrametric Fitting problem.*

This result finds applications in the ℓ_p-Tree Fitting problem, where the goal becomes finding a tree metric that minimizes the same ℓ_p distance to the input pairwise distance matrix. Directly using a standard reduction from fitting tree metrics to fitting ultrametrics [ABF+99; CDK+24], we obtain a constant-factor approximation for $p > 1$, improving the best-known $O((\log n \log\log n)^{1/p})$-approximation [AC11]. See Subsection 1.2 for more prior works on this problem.

Corollary 2. *For any $p > 1$, there exists an $O(p)$-approximation for the ℓ_p-Tree Fitting problem.*

With the same algorithm and an even simpler analysis, we show that the algorithm also improves the current best-known approximation for the ℓ_1-Ultrametric Fitting problem.

Theorem 3. *There exists a 12.091-approximation for the ℓ_1-Ultrametric Fitting problem.*

1.1 Technical overview

Correlation Clustering and the standard LP relaxation. The ℓ_p-Ultrametric Fitting problem has a deep connection with the well-studied Correlation Clustering problem, where we are given a complete graph over n nodes and a $+/-$ label on each edge, and the goal is to find a clustering (i.e., several disjoint cliques that partition the n nodes) such that the number of mistakes (i.e. the number of inter-cluster $+$edges plus the number of intra-cluster $-$edges) is minimized. Correlation Clustering is a special case of ℓ_1-Ultrametric Fitting when the input distance matrix consists of only two different values.

Suppose that $D_1 < \cdots < D_L$ are the distinct values in the input matrix X. Standard arguments [HKM05; AC11] allow us to restrict our attention to output ultrametrics whose distances are all in $\{D_1, \ldots, D_L\}$, by losing only a constant factor on the ratio. An ℓ_p-Ultrametric Fitting instance can then be viewed as L layers of Correlation Clustering instances, where

- The input of layer ℓ consists of positive edges $E^\ell_+ = \{ij : X(i,j) \le D_\ell\}$ and negative edges $E^\ell_- = \{ij : X(i,j) > D_\ell\}$, whose input distances are at most and strictly larger than D_ℓ, respectively.
- The output should be L layers of clustering such that clusters in a lower layer are refinements of clusters in a higher layer. The clustering on the layer ℓ should agree with E^ℓ_+ and E^ℓ_- as much as possible.

Suppose the output ultrametric is U. We can construct a clustering for each layer ℓ as follows: we say i, j are in the same cluster on layer ℓ if $U(i,j) \le D_\ell$. With the ultrametric triangle inequality, this is guaranteed to be a clustering.

The standard integer program of Correlation Clustering considers a variable $x(i,j) \in \{0,1\}$ for each pair of nodes i, j, indicating whether i, j are in different clusters. In other words, $x(i,j) = 1$ corresponds to i and j in different clusters and $x(i,j) = 0$ corresponds to i and j in the same cluster. The only constraint is a triangle inequality $x(i,j) \le x(i,k) + x(k,j)$ for any three nodes i, j, k. In the standard LP relaxation of this integer program, we relax the $\{0,1\}$ integer constraint to $x(i,j) \in [0,1]$, and the triangle inequality constraints imply that a valid LP solution x defines a metric (technically a pseudometric, since 0 distances are allowed). The objective is then $\sum_{ij \in E_+} x(i,j) + \sum_{ij \in E_-} (1 - x(i,j))$, where E_+ and E_- are the sets of $+$edges and $-$edges.

The standard LP relaxation for ℓ_p-Ultrametric Fitting then consists of L copies of the standard LP relaxation for Correlation Clustering. The only additional constraint is to

ensure consistency across different layers: $x_\ell(i,j) \geq x_{\ell+1}(i,j)$ for each layer ℓ and nodes i, j. The objective is slightly more complicated and can be viewed as a relaxation of the p^{th} power of the ℓ_p error objective $||U - X||_p^p$: we can naturally define a distribution $\mu(i,j)$ over $\{D_1, \ldots, D_L\}$ using $x_1(i,j), \ldots, x_L(i,j)$, and let the p^{th} power of the objective be $\sum_{ij} \mathbb{E}_{Y(i,j)\sim\mu(i,j)}[|Y(i,j) - X(i,j)|^p]$. See Section 2 for more details.

Our algorithm. Our algorithm is a simple LP rounding algorithm using the standard LP. We consider a parameter $\delta \in (0, 0.5)$ which is a small constant. At the beginning of the algorithm, we put all species in singleton clusters. Then, we run the following process bottom-up, from the bottom-most layer 1 to the topmost layer L. At each layer ℓ, if there exist two clusters A and B such that the average LP distance $x_\ell(A,B) := \frac{1}{|A|\cdot|B|}\sum_{i\in A, j\in B} x_\ell(i,j)$ is at most δ, we merge them into a single cluster $A \cup B$. We repeat this until no such pair of clusters exists in the layer. The final distance between two species in the output ultrametric is determined by the bottommost layer (i.e. maximum layer ℓ) where they are in the same cluster.[1]

Similar merging algorithms have been used to obtain the constant-factor approximations for the ℓ_1 objective [CDK+24; AKLL25]. Merging algorithms in those papers rely on a pre-processed "clean" LP solution or a pre-clustering, and then maintain a core for each cluster in which the pairwise LP distances are all small. It is very surprising that such a simple merging algorithm (reminiscent of average-linkage hierarchical clustering) improves the approximation ratio for the ℓ_1 objective and gives the first constant-factor any ℓ_p objective.

New ingredients in the analysis. In the analysis, we first consider a δ-merge forest for each layer ℓ, where the vertices are clusters (i.e. subsets) of the n species . The layer-ℓ merge forest consists of all clusters we have seen in our merge process for the first ℓ layers (including the singletons). If two clusters A, B are merged to C, we view A, B as two children of C.

It would not be hard to show by induction that the average intra-cluster LP distance (i.e., $x_\ell(A,A)$) for each vertex A in the layer-ℓ merge forest is small. The analysis is based on a new mathematical property of that could be of independent interest: Let $\phi(x) := \min\{x, 1-x\}$ be the minimum possible LP contribution of an edge with LP distance x. For any two vertices A, B in the merge forest, we can lower bound the minimum possible average LP contribution $\mathbb{E}_{i\sim A, j\sim B}[\phi(x_\ell(i,j))]$ for edges in the biclique $A \times B$, simply by $\Omega(1)\cdot\phi(x_\ell(A,B))$. In other words, the minimum possible average LP contribution from the biclique is $\Omega(1)$, once $x_\ell(A,B)$ is not close to 0 or 1 (especially near the boundary δ of whether A and B should be merged). As this is the reverse direction of Jensen's inequality for the concave function ϕ, we will call this the *approximate inverse Jensen's inequality for* ϕ.

The second new ingredient of the analysis is a new way to decompose the cost of the algorithm and the LP cost: *the biclique decomposition*. Previous papers on the ℓ_1 objective all consider a layer-by-layer decomposition. For example, we decompose the total LP cost into $L-1$ parts, with the ℓ-th part being $D_{\ell+1} - D_\ell$ times the LP cost of x_ℓ in the Correlation Clustering problem (with respect to E_+^ℓ and E_ℓ^-). Then, we upper bound each layer's algorithmic cost by the LP cost. This decomposition has a natural analytical barrier for ℓ_p objectives for $p > 1$: on each layer, we have a weighted Correlation Clustering problem. Unlike the uniform weight Correlation Clustering problem which admits constant factor approximatins, the weighted version is known to be UGC-hard for any constant-factor approximation.

The new biclique decomposition considers each merge as a separate part: if we merge A, B at some time point of the algorithm, we consider all edges in the biclique $A \times B$ as one part. Because

[1] By enforcing $x_L(i,j) \equiv 0$, we can always get a single cluster with all n species at the end of the algorithm.

we finally have a cluster involving all species, every pair is merged at some point; hence, the edges of all such bicliques partition $\binom{[n]}{2}$. More specifically, if we merge A and B on layer ℓ^* (so that $U(i,j) = D_{\ell^*}$ for $ij \in A \times B$), the algorithmic cost and the LP cost in this biclique are as follows:

- The algorithmic cost on biclique: $\sum_{i\in A, j\in B} |D_{\ell^*} - X(i,j)|^p$, the total ℓ_p^p errors on edges of the biclique.
- The LP cost on the biclique: $\sum_{i\in A, j\in B} \mathbb{E}_{Y(i,j)\sim\mu(i,j)}[|Y(i,j) - X(i,j)|^p]$.

To obtain an approximation ratio of Γ, it suffices to show for every part, the algorithm's cost is upper bound by Γ times the LP cost.

Then, let's consider layers below ℓ^*[2], and we consider edges $ij \in A \times B$ whose $X(i,j) \le D_{\ell^*}$. By a telescoping argument, the algorithmic cost on the biclique can be expressed by

$$\sum_{\ell=1}^{\ell^*-1} ((D_{\ell^*} - D_\ell)^p - (D_{\ell^*} - D_{\ell+1})^p) \cdot |E_+^\ell \cap (A \times B)| \tag{1}$$

because species in A and B are in different clusters on these layers and the mistakes are the number of $+$ edges between A and B on such layers. With the previous approximate inverse Jensen's inequality, we are able to upper bound $|E_+^\ell \cap (A \times B)|$ by a constant times the unweighted LP cost on the biclique in layer ℓ:

$$\sum_{i\in A, j\in B} \mathbb{E}_{Y(i,j)\sim\mu(i,j)}[|\mathbb{1}(Y(i,j) \le D_\ell) - \mathbb{1}(X(i,j) \le D_\ell)|].^{3}$$

Putting this upper bound back to Equation (1) and using a similar telescoping argument, Equation (1) can be further upper bounded by (where $(\cdot)_+ = \max\{0, \cdot\}$)

$$O(1) \cdot \sum_{ij\in A\times B} \mathbb{E}_{Y(i,j)\sim\mu(i,j)} \left[\left|(D_{\ell^*} - Y(i,j))_+^p - (D_{\ell^*} - X(i,j))_+^p\right|\right] .$$

With the help of the Minkowski's inequality (Lemma 6), we can upper bound the algorithmic cost by a constant times the LP cost on this biclique $A \times B$.

1.2 Further related work

Fitting ultrametrics with other objectives. The problem of fitting ultrametrics for ℓ_∞ and ℓ_0 norms are well-studied in the literature. Farach et al. [FKW95] gave a polynomial-time algorithm for minimizing the ℓ_∞ objective. The problem of fitting ultrametrics for ℓ_0 norm was introduced later by [CFLM25] and the current best-known ratio is 5 [CG24; AKLL25].

Many other objectives and assumptions have also considered in the literature on fitting ultrametrics, including the maximization version [CFLM25], the outlier deletion version [SWW17], multiplicative distortion version [DPS15] and assuming the input is in high-dimensional Euclidean space [CKL20; CDL21].

[2] layers above ℓ^* use almost the same arguments

[3] This is in fact equal to $\sum_{ij\in E_+^\ell \cap (A\times B)} x_\ell(i,j) + \sum_{ij\in E_-^\ell \cap (A\times B)} 1 - x_\ell(i,j)$, matching the cleaner LP objective of Correlation Clustering on layer ℓ.

Fitting tree metrics. Cavalli-Sforza and Edwards [CE67] introduced the tree fitting problem. If there is a tree metric that fits the data exactly, it can also be found efficiently [WSSB77]. However, if there is no such tree metric, the problem for any ℓ_p objective ($1 \le p \le \infty$) is APX-hard [ABF+99; CGW05]. In addition, Agarwala et al. [ABF+99] and Cohen-Addad et al. [CDK+24] give a general reduction from tree metrics to ultrametrics, only losing a factor of $3 \cdot 2^{(p-1)/p}$. The state-of-art algorithms are then a 3-approximation for ℓ_∞ norm, an $(77.3538 + \varepsilon)$-approximation algorithm for ℓ_1 norm and $O((\log n \log\log n)^{1/p})$-approximations for ℓ_p norms ($1 < p < \infty$). Later, [Kip23] gave a reduction for ℓ_0 norm losing a factor of 6, implying a 30-approximation.

Fitting metrics. Brickell et al. [BDST08] formulated the problem of fitting metrics to a given data with ℓ_p objective for $1 \le p \le \infty$, which can be efficiently solved via linear or convex programs. The ℓ_0 version was introduced later and studied by [GJ17; FRV22; FGR+20; CFLM25], culminating in an $O(\log n)$-approximation algorithm that runs in $O(n^3)$ time.

Correlation clustering. As mentioned before, Correlation Clustering has been extensively studied. Bansal, Blum and Chawla [BBC04] introduced the problem and gave a constant factor approximation. Later, it was improved by [CGW05; ACN08; CMSY15; CLN22; CLLN23; CCL+24; GS26]. The current best-known peer-reviewed algorithm is a $(1.485 + \varepsilon)$-approximation based on rounding a set-based LP whose solution can be efficiently computed via a Sherali-Adams hierarchy [CCL+24]. A recent preprint [GS26] claims a $(1.3865 + \varepsilon)$-approximation. This problem is known to be APX-hard [CGW05].

1.3 Discussion

What we are most excited about is the fact that the results in this work come from an extremely simple algorithm. It was surprising to us that this algorithm manages to obtain the first constant factor approximation results for fitting an ultrametric to given data with ℓ_p error for any $p > 1$ and improve the previous best constant approximation ratio for the well studied ℓ_1 objective. While we have chosen to present the simplest analysis of this algorithm (that we are aware of) in this work, the simplicity of the algorithm naturally lends itself to more sophisticated analyses. We briefly mention improved approximation factors for various ultrametric fitting problems that were obtained for this algorithm in extended interactions with GPT-6 Astra. The complexity of these improved analyses is much higher than the simple analysis we present here. We should state clearly that we do not understand these improved analyses, but we did generate Lean verifications for these results, which gives some degree of confidence in their accuracy. Note however that this complexity is purely in the analysis – the algorithm itself is the simple bottom-up average-linkage hierarchical clustering algorithm applied to the LP solution.

1. For ℓ_1-Ultrametric Fitting, the algorithm achieves an approximation factor at most 7.18134.

2. For ℓ_2-Ultrametric Fitting, the algorithm achieves an approximation factor at most 14.68807.

3. For ℓ_p-Ultrametric Fitting, the algorithm achieves an approximation factor at most $(7.5)^{1/p}/(7.7^{1/p} - (6.7)^{1/p})$.

In Section 5 we show that a lower bound of 6.08155 for the approximation ratio of this algorithm for ℓ_1-Ultrametric Fitting. Thus it appears that the approximation ratio of this approach is in the interval $(6.08155, 7.18134)$.

The analyses in this work run the algorithm for a fixed value of δ, which is the threshold that determines how cluster merging decisions are made. A natural approach to improve the approximation factor would be to try all possible values of δ and output the best clustering or adapt the value of the δ to different layers. While we have not managed to obtain a better upper bound using these ideas, in interactions with GPT-6 Astra, we obtained a construction that shows a lower bound of 5.15134 for the algorithm that tries multiple values of δ and reports the best solution.

2 Preliminaries

In this paper, we use $[n]$ to denote $\{1, 2, \ldots, n\}$. We use $i \sim A$ to denote that i is sampled uniformly at random in a set A.

Problem definition and a basic reduction. In the ℓ_p-Ultrametric Fitting problem, we are given a set of n species, and a pairwise distance matrix $X : \binom{[n]}{2} \to \mathbb{R}_{>0}$ between the species. The goal is to find an ultrametric $U : \binom{[n]}{2} \to \mathbb{R}_{>0}$ that minimizes the ℓ_p-distance to the input distance matrix X, i.e.,

$$\|U - X\|_p = \left(\sum_{i,j \in \binom{[n]}{2}} |U(i,j) - X(i,j)|^p \right)^{1/p}.$$

Throughout the paper, we assume that $p \geq 1$.

The ℓ_1-Ultrametric Fitting problem is a generalization of the well studied Correlation Clustering problem.

We will use $D_1 < D_2 < \cdots < D_L$ to denote the L distinct distance in X. We show that we can reduce the problem to the case where the output U only takes values in $\{D_1, \ldots, D_L\}$, but only losing a factor of 2 on the approximation ratio.

Lemma 4 ([HKM05])**.** *For any $X : \binom{[n]}{2} \to \mathbb{R}_{>0}$ and $p \geq 1$, there exists an ultrametric $U : \binom{[n]}{2} \to \mathbb{R}_{>0}$ such that $U(i,j) \in \{D_1, \ldots, D_L\}$ for all $i, j \in \binom{[n]}{2}$ and $\|U - X\|_p \leq 2 \cdot \min_{U'} \|U' - X\|_p$, where the minimum is taken over all ultrametrics U'.*

In particular, if $p = 1$, there exists one such ultrametric U with $\|U - X\|_1 = \min_{U'} \|U' - X\|_1$.

In the rest of the paper, we will focus on algorithm whose the output ultrametric U only takes values in $\{D_1, \ldots, D_L\}$.

LP formulation. We consider the standard LP formulation for the ℓ_p-Ultrametric Fitting problem. In the standard integer program that characterize the problem, we have a variable $x_\ell(i,j)$ for each layer $\ell \in [L]$ and each pair of species $i, j \in \binom{[n]}{2}$, which indicates whether the distance between species i and j is strictly greater than D_ℓ. Naturally, by definition, we have a constraint that $x_\ell(i,j) \geq x_{\ell+1}(i,j)$ for all $\ell \in [L-1]$ and $i, j \in [n]$, and $x_L(i,j) = 0$ for all $i, j \in \binom{[n]}{2}$. Because ultrametrics satify a stronger triangle inequality $U(i,j) \leq \max\{U(i,k), U(k,j)\}$, we have the following triangle inequality constraints for the variables: $x_\ell(i,j) \leq x_\ell(i,k) + x_\ell(k,j)$ for all $\ell \in [L]$ and $i, j, k \in [n]$. For simplicity, we will define $x_0(i,j) = 1$ for all $i, j \in \binom{[n]}{2}$. Putting these together, we have the following LP relaxation for the ℓ_p-Ultrametric Fitting problem, where

the p-th power of the ℓ_p-distance is minimized.

$$\begin{aligned}
\text{minimize} \quad & \sum_{i,j\in\binom{[n]}{2}} \mathsf{lp\text{-}cost}_p(x,i,j) && (\ell_p\text{-ultrametric-LP})\\
\text{subject to} \quad & x_\ell(i,j) \le x_\ell(i,k) + x_\ell(k,j) && \forall \ell \in [L], i,j,k \in [n] \quad (2)\\
& x_L(i,j) = 0 && \forall i,j \in [n] \quad (3)\\
& x_\ell(i,j) \ge x_{\ell+1}(i,j) && \forall \ell \in [L-1], i,j \in [n] \quad (4)\\
& x_\ell(i,j) \in [0,1] && \forall \ell \in [L], i,j \in [n] \quad (5)
\end{aligned}$$

where we define the LP cost of each pair of species i, j as

$$\mathsf{lp\text{-}cost}_p(x,i,j) = \sum_{\ell=1}^{L} |D_\ell - X(i,j)|^p \cdot (x_{\ell-1}(i,j) - x_\ell(i,j)) \ .$$

The cost function $\mathsf{lp\text{-}cost}_p(x,i,j)$ can be interpreted as the expectation of $|Y(i,j) - X(i,j)|^p$ where $Y(i,j)$ is a random variable that takes value D_ℓ with probability $x_{\ell-1}(i,j) - x_\ell(i,j)$ for each $\ell \in [L]$. Further, we define $\mathsf{lp\text{-}cost}_p(x) := \sum_{i,j\in\binom{[n]}{2}} \mathsf{lp\text{-}cost}_p(x,i,j)$ to be the LP cost for solution x.

Distance between two sets. Naturally, the LP solution x_ℓ on each layer ℓ defines a metric on the n species. Next, we give a useful definition of the LP distance between two non-empty sets of species based on any metric d. For a metric d and two sets A, B, we define the distance between the sets as the average distance between all pairs of elements in the two sets, i.e.,

$$d(A,B) = \frac{1}{|A||B|} \cdot \sum_{i\in A, j\in B} d(i,j).$$

Lemma 5. *For any three sets A, B, C, we have $d(A,C) \le d(A,B) + d(B,C)$.*

Proof. By the triangle inequality for the metric d, we have

$$\begin{aligned}
d(A,C) &= \frac{1}{|A||C|} \cdot \sum_{i\in A, k\in C} d(i,k)\\
&\le \frac{1}{|A||C|} \cdot \sum_{i\in A, j\in B, k\in C} \frac{d(i,j) + d(j,k)}{|B|}\\
&= \frac{1}{|A||B|} \cdot \sum_{i\in A, j\in B} d(i,j) + \frac{1}{|B||C|} \cdot \sum_{j\in B, k\in C} d(j,k)\\
&= d(A,B) + d(B,C).
\end{aligned}$$

□

Minkowski's inequality. We will use Minkowski's inequality extensively in the proof.

Lemma 6 (Minkowski's inequality)**.** *For any random variable X, Y supported over $\mathbb{R}$ and any $1 \le p < \infty$, we have*

$$\mathbb{E}[|X+Y|^p]^{1/p} \le \mathbb{E}[|X|^p]^{1/p} + \mathbb{E}[|Y|^p]^{1/p} \ .$$

Algorithm 1 (Merge(δ)).

Initialization: LP solution $\{x_\ell(i,j)\}_{\ell\in[L],i,j\in[n]}$; initial clustering $\mathcal{C} = \{\{i\} : i \in [n]\}$; input distances $D_1, \ldots, D_L$.

For each $\ell \in [L]$,

1. While there exists a pair of clusters $A, B \in \mathcal{C}$ such that $x_\ell(A,B) \le \delta$, merge them into a single cluster $A \cup B$ and update $\mathcal{C} \leftarrow (\mathcal{C} \setminus \{A,B\}) \cup \{A \cup B\}$.
2. Let $\mathcal{C}^\ell = \mathcal{C}$.

Define U as follows: for each $i,j \in \binom{[n]}{2}$, we set let $\ell^*(i,j)$ be the minimum index $\ell \in [L]$ such that i,j are in the same cluster in $\mathcal{C}^\ell$ and $U(i,j) = D_{\ell^*(i,j)}$.

Output: U.

3 Algorithm

Suppose that $\{x_\ell(i,j)\}_{\ell\in[L],i,j\in[n]}$ is an optimal solution for the ultrametric LP (ℓ_p-ultrametric-LP). For each $\ell \in [L]$, we define the metric $([n], x_\ell)$ to be the metric naturally induced by the solution x_ℓ[4]. With this LP solution, the algorithm works in a bottom-up manner and is parameterized by a constant $\delta \in (0, 0.5)$. At the beginning, we put all species in their singletons. Then, we go from the bottommost layer to the topmost layer, merging clusters based on the LP solution at each layer. At each layer ℓ, if there exist two clusters A and B such that $x_\ell(A,B) \le \delta$, we merge them into a single cluster $A \cup B$. We repeat this until no such pair exists in the layer. The final distance between two species in the ultrametric is determined by the layer at which they are first in the same cluster. See Algorithm 1 for the pseudocode of the algorithm.

Since $x_L(i,j) = 0$ (Equation (3)), all points belong to one cluster by the end of layer L. Hence $\ell^*(i,j)$ exists for every $ij \in \binom{[n]}{2}$, and U is well-defined. Further, since i,j and j,k are in the same cluster at layer $\max\{\ell^*(i,j), \ell^*(j,k)\}$, i,k should also appear in the same cluster. We have $\ell^*(i,k) \le \max\{\ell^*(i,j), \ell^*(j,k)\}$ and thus the output U is an ultrametric.

4 Analysis

4.1 A perspective of merge forest

We first introduce the notion of a δ-merge forest, which consists of binary trees that can be used to represent the merging process of the algorithm. In this subsection, we will consider an arbitrary metric $([n], d)$.

Definition 7 (δ-merge forest)**.** *A δ-merge forest with respect to the metric $([n], d)$ is a forest consisting of binary trees satisfying the following properties.*

- *Each node is represented by a non-empty subset of $[n]$.*
- *The leaf nodes are different singletons, i.e., each leaf node corresponds to a distinct element in $[n]$.*

[4]For convention, we enforce $x_\ell(i,i) = 0$ for all $i \in [n]$ here.

- *For any non-leaf node A, it has exactly two children A_1, A_2 such that $A = A_1 \cup A_2$, $A_1 \cap A_2 = \emptyset$ and $d(A_1, A_2) \le \delta$.*

We present two simple claims on the intra-cluster LP distance for each node in the δ-merge forest.

Claim 8. *For each node A in a δ-merge forest, we have $d(A, A) \le \delta$.*

Proof. We prove this claim by induction on the size of A. The base case is when $|A| = 1$. In this case, $d(i, j) = 0$ for the only element $i = j \in A$, and the claim holds trivially.

For the induction step, let A_1 and A_2 be the children of a non-leaf node A in the δ-merge forest. By the induction hypothesis, we have $d(A_1, A_1) \le \delta$ and $d(A_2, A_2) \le \delta$. Let $t = |A_1|/|A|$, we have

$$\begin{aligned} d(A, A) &= t^2 d(A_1, A_1) + (1-t)^2 d(A_2, A_2) + 2t(1-t) d(A_1, A_2) \\ &\le t^2\delta + (1-t)^2\delta + 2t(1-t)\delta \le \delta. \end{aligned}$$

□

This claim can generalize to the interpolation of the two children of a non-leaf node in the δ-merge forest, which is defined as follows.

Definition 9. *For each non-leaf node A in a δ-merge forest with children A_1 and A_2, we define its interpolation, parameterized by $t \in [0, 1]$, as a distribution $\mu_t(A)$ over the elements of A such that it gives a uniform element from A_1 with probability t and a uniform element from A_2 with probability $1 - t$. That is, if $X \sim \mu_t(A)$, we have*

$$\forall i \in A, \quad \Pr[X = i] = \begin{cases} t/|A_1| & \text{if } i \in A_1, \\ (1-t)/|A_2| & \text{if } i \in A_2. \end{cases}$$

Claim 10. *For each non-leaf node A in a δ-merge forest and each $t \in [0, 1]$, we have*

$$\mathbb{E}_{i,j \sim \mu_t(A)}[d(i, j)] \le \delta.$$

Proof. Let A_1 and A_2 be the children of A in the δ-merge forest. By Claim 8, we have $d(A_1, A_1) \le \delta$ and $d(A_2, A_2) \le \delta$. By definition of the δ-merge forest, we have $d(A_1, A_2) \le \delta$. Therefore, by the definition of the interpolation, we have

$$\begin{aligned} \mathbb{E}_{i,j \sim \mu_t(A)}[d(i, j)] &= t^2 \mathbb{E}_{i,j \sim A_1}[d(i, j)] + (1-t)^2 \mathbb{E}_{i,j \sim A_2}[d(i, j)] + 2t(1-t)\mathbb{E}_{i \sim A_1, j \sim A_2}[d(i, j)] \\ &= t^2 d(A_1, A_1) + (1-t)^2 d(A_2, A_2) + 2t(1-t) d(A_1, A_2) \\ &\le t^2\delta + (1-t)^2\delta + 2t(1-t)\delta \le \delta. \end{aligned}$$

□

4.2 An approximate inverse Jensen's inequality

Let $\phi(x) = \min\{x, 1 - x\}$. Consider any metric $([n], d)$ such that $d(\cdot, \cdot) \in [0, 1]$. On a fixed layer, an edge with LP distance x contributes either x or $1 - x$ to the unweighted (disagreement) cost, according to its label. Hence, $\phi(x)$ is a lower bound on that cost, independently of its label. It is clear that $\phi(x)$ is a concave function on $[0, 1]$ and hence Jensen's inequality gives $\phi(\mathbb{E}[X]) \ge \mathbb{E}[\phi(X)]$ for any random variable X taking values in $[0, 1]$. Next, we establish an approximate inverse Jensen's inequality for the function ϕ, taking expectation over the distance of a random edge in a biclique $A \times B$ for any two nodes A, B in the merge forest. The reason that this inequality is useful in upper bounding the cost of our algorithm is that $\mathbb{E}_{i \sim A, j \sim B}[\phi(d(i, j))]$ is a simple lower bound the average LP cost for edges in $A \times B$. The lemma can also give a lower bound that is reasonably large when $d(A, B)$ is not too close to 0 or 1.

Lemma 11. *Let A, B be two nodes in the δ-merge forest. Then, we have*

$$\mathbb{E}_{i\sim A, j\sim B}[\phi(d(i,j))] \geq (1-2\delta)^2 \phi(d(A,B)) \ .$$

To prove Lemma 11, we consider the following lemma, which we will use twice (and this is why we have a factor of $(1-2\delta)^2$ on the RHS in Lemma 11).

Lemma 12. *Let A be any node in the δ-merge forest. Let $f : A \to [0,1]$ be any function that satisfies $|f(i) - f(j)| \leq d(i,j)$ for all $i, j \in A$. Then, we have*

$$\mathbb{E}_{i\in A}[\phi(f(i))] \geq (1-2\delta) \cdot \phi(\mathbb{E}_{i\in A}[f(i)]) \ .$$

Proof. We prove this lemma by induction. The base case is when A is a leaf node. In this case, the LHS equals to $\phi(f(i))$ and the RHS equals to $(1-2\delta)\phi(f(i))$ for the single element $i \in A$.

Now, assume that the lemma holds for all descendants of a non-leaf node A. Let A_1 and A_2 be the children of A in the δ-merge forest. By our induction hypothesis, we have

$$\mathbb{E}_{i\sim\mu_1(A)}[\phi(f(i))] = \mathbb{E}_{i\in A_1}[\phi(f(i))] \geq (1-2\delta)\cdot\phi(\mathbb{E}_{i\in A_1}[f(i)]) = (1-2\delta)\cdot\phi(\mathbb{E}_{i\sim\mu_1(A)}[f(i)]) \ , \quad (6)$$

$$\mathbb{E}_{i\sim\mu_0(A)}[\phi(f(i))] = \mathbb{E}_{i\in A_2}[\phi(f(i))] \geq (1-2\delta)\cdot\phi(\mathbb{E}_{i\in A_2}[f(i)]) = (1-2\delta)\cdot\phi(\mathbb{E}_{i\sim\mu_0(A)}[f(i)]) \ . \quad (7)$$

Let $t^* = \frac{|A_1|}{|A|}$. Our goal is to prove that

$$\mathbb{E}_{i\sim\mu_{t^*}(A)}[\phi(f(i))] \geq (1-2\delta)\cdot\phi(\mathbb{E}_{i\sim\mu_{t^*}(A)}[f(i)]) \ .$$

Note that $\phi(x)$ is a linear function on $[0, 0.5]$ (or on $[0.5, 1]$). If $\mathbb{E}_{i\sim\mu_0(A)}[f(i)]$ and $\mathbb{E}_{i\sim\mu_1(A)}[f(i)]$ are both less than or equal to 0.5, or both greater than or equal to 0.5, then we can directly average the two inequalities (6) and (7) to get our goal. Next, without loss of generality, we consider the case where $\mathbb{E}_{i\sim\mu_0(A)}[f(i)] < 0.5 < \mathbb{E}_{i\sim\mu_1(A)}[f(i)]$. Let $t \in [0,1]$ be the real number such that $\mathbb{E}_{i\sim\mu_t(A)}[f(i)] = 0.5$. Note that $\phi(x) = 0.5 - |x - 0.5|$.

$$\begin{aligned}
\mathbb{E}_{i\sim\mu_t(A)}[\phi(f(i))] &= \mathbb{E}_{i\sim\mu_t(A)}\left[0.5 - |f(i) - 0.5|\right] \\
&= 0.5 - \mathbb{E}_{i\sim\mu_t(A)}\left[\left|f(i) - \mathbb{E}_{j\sim\mu_t(A)}[f(j)]\right|\right] \\
&\geq 0.5 - \mathbb{E}_{i,j\sim\mu_t(A)}\left[|f(i) - f(j)|\right] && (|a-x| \text{ is convex for } x) \\
&\geq 0.5 - \mathbb{E}_{i,j\sim\mu_t(A)}\left[d(i,j)\right] && (\text{assumption on } f) \\
&\geq 0.5 - \delta = (1-2\delta)\cdot\phi(\mathbb{E}_{i\sim\mu_t(A)}[f(i)]) && (\text{by Claim 10})
\end{aligned}$$

Note that $\phi(x) = x = 1 - x$ when $x = 0.5$. If $t^* \in [t, 1]$, then we can average the above inequality with (6) to get our goal. Otherwise, we can average the above inequality with (7) to get our goal. □

Proof of Lemma 11. First, let us consider $f_i : B \to [0,1]$ defined by $f_i(j) = d(i,j)$ for any $i \in A$. Because of the triangle inequality, we have $|f_i(j) - f_i(j')| \leq d(j,j')$ for any $j, j' \in B$. Therefore, we can apply the result from Lemma 12 to the function f_i for any $i \in A$ to get the following inequality:

$$\mathbb{E}_{i\sim A, j\sim B}[\phi(d(i,j))] \geq (1-2\delta)\mathbb{E}_{i\sim A}[\phi(\mathbb{E}_{j\sim B}[d(i,j)])] = (1-2\delta)\mathbb{E}_{i\sim A}[\phi(d(i,B))]$$

Next, consider $g : A \to [0,1]$ defined by $g(i) = d(i,B)$. Because of the triangle inequality for the set (Lemma 5), we have $|g(i) - g(i')| \leq |d(i,B) - d(i',B)| \leq d(i,i')$. Therefore, we can apply the result from Lemma 12 to the function g to get the following inequality:

$$\mathbb{E}_{i\sim A}[\phi(d(i,B))] \geq (1-2\delta)\phi(\mathbb{E}_{i\sim A}[d(i,B)]) = (1-2\delta)\phi(d(A,B)) \ .$$

Combining the two inequalities, we get

$$\mathbb{E}_{i\sim A, j\sim B}[\phi(d(i,j))] \geq (1-2\delta)^2\phi(d(A,B)) \ .$$ □

4.3 Biclique decomposition of the algorithm's cost

In this subsection, we show that how we decompose the algorithmic cost of Algorithm 1 and the LP cost of x into disjoint bicliques (whose edges are pairwise disjoint). We will analyze each biclique separately in the next subsection to obtain the constant-factor approximation for any ℓ_p objective.

First, we define the δ-merge forest T_ℓ on each layer ℓ. T_ℓ will be built on $T_{\ell-1}$, where we assume that there are only singletons in T_0. It is easy to see that $T_{\ell-1}$ is a δ-merge forest on layer ℓ, because the LP constraint (4). If at layer ℓ, we merge A and B, we build a new node $A \cup B$, labeled by ℓ, and let A, B be the two children of it. We shall remark that The toots of T_ℓ are precisely the clusters in $\mathcal{C}^\ell$. We shall also remark that because of the LP constraint Equation (3), we have a node corresponding to the cluster $[n]$ involving all species in T_L.

Then, the decomposition follows: for each node $A \cup B$ labeled by ℓ^* (denoted as node $(A \cup B, \ell^*)$ from now on) in T_L, we consider the cost of the algorithm for this node as the sum of algorithmic costs on all edges in the biclique $A \times B$. That is, for node $(A \cup B, \ell^*)$, the algorithmic cost is

$$\mathsf{cost}(A, B, \ell^*) = \sum_{i \in A, j \in B} |D_{\ell^*} - X(i,j)|^p \ .$$

On the other hand, we also decompose the total LP cost (ℓ_p-ultrametric-LP) into the sum of LP costs on all edges in the biclique $A \times B$ for each node $(A \cup B, \ell^*)$, which is

$$\mathsf{lp}(A \times B, x) = \sum_{i \in A, j \in B} \mathsf{lp\text{-}cost}_p(x, i, j) \ .$$

To show that Algorithm 1 is a Γ-approximation, it suffices to show that for each node $(A \cup B, \ell^*)$, we have

$$(\mathsf{cost}(A, B, \ell^*))^{1/p} \le \Gamma \cdot (\mathsf{lp}(A \times B, x))^{1/p} \ . \tag{8}$$

We will formalize why this gives a Γ-approximation at the end of the next subsection (the proof of Theorem 1).

4.4 Constant-factor approximation for ℓ_p-objective

In this subsection, we show Equation (8) for $\Gamma = O(p)$ for any fixed $p \ge 1$, and then prove Theorem 1.

Lemma 13. *Let $\Gamma_0 = 1 + \frac{1-\delta}{\delta(1-2\delta)^2}$. For each node $(A \cup B, \ell^*)$ in T_L, we have*

$$(\mathsf{cost}(A, B, \ell^*))^{1/p} \le \Gamma \cdot (\mathsf{lp}(A \times B, x))^{1/p} \ ,$$

where $\Gamma = \frac{2}{((\Gamma_0+1)/\Gamma_0)^{1/p}-1}$.

For each $\ell \in [L]$ and two nodes A, B in T_L, we define two sets $E^\ell_+(A,B)$ and $E^\ell_-(A,B)$ that denote the $+$edges and $-$edges (following the convention of Correlation Clustering) on layer ℓ in the biclique $A \times B$, respectively.

$$E^\ell_+(A,B) = \{(i,j) \in A \times B : X(i,j) \le D_\ell\} \ , \quad \text{and} \quad E^\ell_-(A,B) = \{(i,j) \in A \times B : X(i,j) > D_\ell\} \ .$$

Further, for each layer $\ell \in [L]$, we consider the unweighted LP cost on layer ℓ as

$$\mathsf{err}_\ell(A,B) = \sum_{ij \in E^\ell_+(A,B)} x_\ell(i,j) + \sum_{ij \in E^\ell_-(A,B)} (1 - x_\ell(i,j)) \ .$$

With Lemma 11, we can prove the following lemma.

Lemma 14. *For every $\ell \in [L]$, $\delta \in (0, 0.5)$ and any two nodes A, B in T_ℓ such that $A \cap B = \emptyset$, we have*

- *If $x_\ell(A, B) \le \delta$, then we have $|E^\ell_-(A, B)| \le (1 + \frac{1}{(1-2\delta)^2})\mathsf{err}_\ell(A, B)$.*
- *If $x_\ell(A, B) > \delta$, then we have $|E^\ell_+(A, B)| \le (1 + \frac{1-\delta}{\delta(1-2\delta)^2})\mathsf{err}_\ell(A, B)$.*

Proof. For the first bullet, we have

$$\begin{aligned}
|E^\ell_-(A,B)| = \sum_{ij \in E^\ell_-(A,B)} 1 &\le \mathsf{err}_\ell(A,B) + \sum_{ij \in A\times B} x_\ell(i,j) \\
&= \mathsf{err}_\ell(A,B) + |A| \cdot |B| \cdot \phi(x_\ell(A,B)) && (x_\ell(A,B) \le \delta \le 1/2) \\
&\le \mathsf{err}_\ell(A,B) + \frac{1}{(1-2\delta)^2} \sum_{ij \in A\times B} \phi(x_\ell(i,j)) && \text{(Lemma 11)} \\
&\le \left(1 + \frac{1}{(1-2\delta)^2}\right) \mathsf{err}_\ell(A,B) \ .
\end{aligned}$$

For the second bullet, we have

$$\begin{aligned}
|E^\ell_+(A,B)| = \sum_{ij \in E^\ell_+(A,B)} 1 &\le \mathsf{err}_\ell(A,B) + \sum_{ij \in A\times B} 1 - x_\ell(i,j) \\
&= \mathsf{err}_\ell(A,B) + |A| \cdot |B| \cdot (1 - x_\ell(A,B)) \\
&\le \mathsf{err}_\ell(A,B) + |A| \cdot |B| \cdot \frac{1-\delta}{\delta} \cdot \phi(x_\ell(A,B)) \ \ (x_\ell(A,B) > \delta \text{ and } \delta < 0.5) \\
&\le \mathsf{err}_\ell(A,B) + \frac{1-\delta}{\delta(1-2\delta)^2} \sum_{ij \in A\times B} \phi(x_\ell(i,j)) && \text{(Lemma 11)} \\
&\le \left(1 + \frac{1-\delta}{\delta(1-2\delta)^2}\right) \mathsf{err}_\ell(A,B) \ ,
\end{aligned}$$

where the third line follows from a simple discussion on whether $x_\ell(A, B) \le 0.5$ or $x_\ell(A, B) > 0.5$. Hence, we have proved the lemma. □

We let $\mathsf{idx}(i, j)$ be the index of the layer $X(i, j)$ belongs to (i.e., $X(i, j) = D_{\mathsf{idx}(i,j)}$). For each $ij \in A \times B$, let $Y(i, j)$ be a random variable for the ideal output ultrametric that takes value D_ℓ with probability $x_{\ell-1}(i, j) - x_\ell(i, j)$ for each $\ell \in [L]$. Then, we can rewrite the LP cost on the biclique $A \times B$ as follows.

$$\mathsf{lp}(A \times B, x) = \sum_{ij \in A\times B} \mathbb{E}[|Y(i,j) - X(i,j)|^p]$$

Further, we can rewrite the unweighted LP cost on layer ℓ as follows.

$$\begin{aligned}
\mathsf{err}_\ell(A,B) &= \sum_{ij \in E^\ell_+(A,B)} \Pr[Y(i,j) > D_\ell] + \sum_{ij \in E^\ell_-(A,B)} (1 - \Pr[Y(i,j) > D_\ell]) \\
&= \sum_{ij \in A\times B} \mathbb{E}[|\mathbb{1}(X(i,j) > D_\ell) - \mathbb{1}(Y(i,j) > D_\ell)|] && (9) \\
&= \sum_{ij \in A\times B} \mathbb{E}[|\mathbb{1}(X(i,j) \le D_\ell) - \mathbb{1}(Y(i,j) \le D_\ell)|] \ . && (10)
\end{aligned}$$

Proof of Lemma 13. In this proof, we define $(x)_+ = \max\{0, x\}$ for any $x \in \mathbb{R}$. With this notation, we can rewrite the cost of the algorithm on the biclique $A \times B$ as follows.

$$(\mathsf{cost}(A, B, \ell^*))^{1/p} = \left(\sum_{ij \in A \times B} |D_{\ell^*} - X(i,j)|^p \right)^{1/p} \tag{11}$$

$$= \left(\sum_{ij \in A \times B} (D_{\ell^*} - X(i,j))_+^p + \sum_{ij \in A \times B} (X(i,j) - D_{\ell^*})_+^p \right)^{1/p}$$

$$\leq \left(\sum_{ij \in A \times B} (D_{\ell^*} - X(i,j))_+^p \right)^{1/p} + \left(\sum_{ij \in A \times B} (X(i,j) - D_{\ell^*})_+^p \right)^{1/p} \tag{12}$$

where the inequality follows from the Minkowski's inequality. Next, we will upper bound the two terms on the right-hand side separately.

Recall our definitions of $E_+^\ell(A, B)$ and $E_-^\ell(A, B)$. We provide an equivalent expression for the two terms in the above equation, using $\mathsf{idx}(\cdot)$.

$$E_+^\ell(A, B) = \{(i, j) \in A \times B : \mathsf{idx}(i, j) \leq \ell\}, \quad \text{and} \quad E_-^\ell(A, B) = \{(i, j) \in A \times B : \mathsf{idx}(i, j) > \ell\}.$$

For the first term on the RHS of Equation (12), we have

$$\begin{aligned}
\sum_{ij \in A \times B} (D_{\ell^*} - X(i,j))_+^p &= \sum_{ij \in A \times B} (D_{\ell^*} - D_{\mathsf{idx}(i,j)})_+^p \\
&= \sum_{ij \in A \times B} \sum_{\ell = \mathsf{idx}(i,j)}^{\ell^* - 1} ((D_{\ell^*} - D_\ell)^p - (D_{\ell^*} - D_{\ell+1})^p) \\
&= \sum_{\ell=1}^{\ell^* - 1} ((D_{\ell^*} - D_\ell)^p - (D_{\ell^*} - D_{\ell+1})^p) \cdot |E_+^\ell(A, B)| \ .
\end{aligned}$$

Consider any $\ell \in [\ell^* - 1]$. Let $A_1, A_2, \ldots, A_r$ be the root nodes in T_ℓ that are descendants of A (including itself, the same below) in T_{ℓ^*}, and let $B_1, B_2, \ldots, B_s$ be the root nodes in T_ℓ that are descendants of B in T_{ℓ^*}. Hence, they are the clusters appearing in $\mathcal{C}^\ell$. Because none of the pairs (A_i, B_j) are eligible to merge at layer ℓ, we have $x_\ell(A_i, B_j) > \delta$ for all $i \in [r], j \in [s]$. Applying Lemma 14, we have

$$|E_+^\ell(A, B)| = \sum_{i=1}^r \sum_{j=1}^s |E_+^\ell(A_i, B_j)| \leq \sum_{i=1}^r \sum_{j=1}^s \Gamma_0 \cdot \mathsf{err}_\ell(A_i, B_j) = \Gamma_0 \cdot \mathsf{err}_\ell(A, B) \ .$$

Therefore, we have

$$\begin{aligned}
&\sum_{ij\in A\times B}(D_{\ell^*}-X(i,j))_+^p \\
&\quad\le \sum_{\ell=1}^{\ell^*-1}\left((D_{\ell^*}-D_\ell)^p-(D_{\ell^*}-D_{\ell+1})^p\right)\cdot\Gamma_0\cdot \mathsf{err}_\ell(A,B) \\
&\quad= \Gamma_0\sum_{\ell=1}^{\ell^*-1}\left((D_{\ell^*}-D_\ell)^p-(D_{\ell^*}-D_{\ell+1})^p\right)\cdot\sum_{ij\in A\times B}\mathbb{E}[|\mathbb{1}(X(i,j)\le D_\ell)-\mathbb{1}(Y(i,j)\le D_\ell)|] \\
&\qquad\qquad\qquad

This gives an upper bound of $\frac{\Gamma}{2} \cdot (\mathsf{lp}(A \times B, x))^{1/p}$ for the first term on the RHS of Equation (12).

For the second term on the RHS of Equation (12), our proof is similar to the first term. The main difference is that we need to use the property $x_\ell(A, B) \le \delta$ for each $\ell \ge \ell^*$ of Algorithm 1. We defer the details to Appendix A. □

Finally, we can establish the constant-factor approximation for ℓ_p-Ultrametric Fitting.

Proof of Theorem 1. We first note that $\Gamma = \frac{2}{(1+1/\Gamma_0)^{1/p}-1} = O(p)$ for any constant $\delta \in (0, 0.5)$.

Because of Lemma 13, for every node $(A \cup B, \ell^*)$, we have

$$\begin{aligned} &\mathsf{cost}(A, B, \ell^*) \le \Gamma^p \cdot \mathsf{lp}(A \times B, x) \\ &\iff \sum_{i \in A, j \in B} |D_{\ell^*} - X(i,j)|^p \le \Gamma^p \cdot \sum_{i \in A, j \in B} \mathsf{lp\text{-}cost}_p(x, i, j) \ . \end{aligned} \tag{13}$$

Consider the final δ-merge forest T_L. We have shown that there is a node corresponding to $[n]$ in the tree.

We say a pair i, j is covered by a node corresponding to a cluster $A \cup B$ if $ij \in A \times B$. Next, we show at the time we create a node corresponding to A, all pairs $\binom{A}{2}$ are covered exactly once by some node in the tree. The base case is simple, when there are only singletons and no pair is in consideration. When we create A by merging A_1 and A_2, pairs in $\binom{A_1}{2}$ and $\binom{A_2}{2}$ are already covered once by an existing node in the tree. The new node A covers all pairs in $A_1 \times A_2$. Then, all pairs in $\binom{A}{2}$ are covered exactly once after creating A.

Let U be the output ultrametric of Algorithm 1, which gives $U(i,j) = D_{\ell^*}$ for each pair $ij \in A \times B$ covered by a node $(A \cup B, \ell^*)$. Therefore, by summing (13) over all node and take the p-th root on each side, we get

$$\left(\sum_{i,j \in \binom{[n]}{2}} |U(i,j) - X(i,j)|^p \right)^{1/p} \le \Gamma \cdot \left(\sum_{i,j \in \binom{[n]}{2}} \mathsf{lp\text{-}cost}_p(x, i, j) \right)^{1/p},$$

which implies that the algorithm is a 2Γ-approximation for the ℓ_p objective, together with Lemma 4. Finally, we shall remark that the above proof also works for $p = 1$, by concluding a Γ-approximation instead. □

4.5 The improved 12.091-approximation for the ℓ_1 objective

Readers familiar with the connection between ℓ_1-Ultrametric Fitting and the Hierarchical Correlation Clustering problem (used in previous works [CDK+24; AKLL25]) may have realized that Lemma 14 implies a $(1 + \frac{1-\delta}{\delta(1-2\delta)^2})$-approximation for ℓ_1-Ultrametric Fitting. In this subsection, we will present a proof for this (i.e., Theorem 3). We will use all the notations from the previous subsection.

Proof of Theorem 3. Let $\delta = \frac{3-\sqrt{5}}{4}$ and $\Gamma := 1 + \frac{1-\delta}{\delta(1-2\delta)^2} < 12.091$.

For each node $(A \cup B, \ell^*)$ in T_L, the algorithmic cost

$$\begin{aligned}
\mathsf{cost}(A, B, \ell^*) &= \sum_{i \in A, j \in B} |D_{\ell^*} - X(i,j)| \\
&= \sum_{ij \in E_+^{\ell^*}(A,B)} (D_{\ell^*} - X(i,j)) + \sum_{ij \in E_-^{\ell^*}(A,B)} (X(i,j) - D_{\ell^*}) \\
&= \sum_{ij \in E_+^{\ell^*}(A,B)} \sum_{\ell = \mathsf{idx}(i,j)}^{\ell^*-1} (D_{\ell+1} - D_\ell) + \sum_{ij \in E_-^{\ell^*}(A,B)} \sum_{\ell=\ell^*}^{\mathsf{idx}(i,j)-1} (D_{\ell+1} - D_\ell) \\
&= \sum_{\ell=1}^{\ell^*-1} (D_{\ell+1} - D_\ell) \cdot |E_+^\ell(A,B)| + \sum_{\ell=\ell^*}^{L-1} (D_{\ell+1} - D_\ell) \cdot |E_-^\ell(A,B)|
\end{aligned}$$

We note that we have proved that $|E_+^\ell(A,B)| \le \Gamma \cdot \mathsf{err}_\ell(A,B)$ for $\ell < \ell^*$ in the proof of Lemma 13. On the other hand, $x_\ell(A,B) \le \delta$ for $\ell \ge \ell^*$ is clear because we merge A, B in layer ℓ^*. Applying Lemma 14, we have

$$\begin{aligned}
\mathsf{cost}(A, B, \ell^*) &\le \Gamma \cdot \sum_{\ell=1}^{L-1} (D_{\ell+1} - D_\ell) \mathsf{err}_\ell(A,B) \\
&= \Gamma \cdot \sum_{\ell=1}^{L-1} (D_{\ell+1} - D_\ell) \cdot \left(\sum_{ij \in E_+^\ell(A,B)} x_\ell(i,j) + \sum_{ij \in E_-^\ell(A,B)} (1 - x_\ell(i,j)) \right) \\
&= \Gamma \cdot \sum_{ij \in A \times B} \left(\sum_{\ell = \mathsf{idx}(i,j)}^{L-1} (D_{\ell+1} - D_\ell) x_\ell(i,j) + \sum_{\ell=1}^{\mathsf{idx}(i,j)-1} (D_{\ell+1} - D_\ell)(1 - x_\ell(i,j)) \right) \\
&= \Gamma \cdot \sum_{ij \in A \times B} \left(\sum_{\ell = \mathsf{idx}(i,j)}^{L-1} (D_{\ell+1} - D_{\mathsf{idx}(i,j)})(x_\ell(i,j) - x_{\ell+1}(i,j)) \qquad (x_L(i,j) = 0) \right. \\
&\qquad \left. + \sum_{\ell=1}^{\mathsf{idx}(i,j)-1} (D_{\mathsf{idx}(i,j)} - D_\ell)(x_{\ell-1}(i,j) - x_\ell(i,j)) \right) \qquad (x_0(i,j) = 1) \\
&= \Gamma \cdot \mathsf{lp}(A \times B, x).
\end{aligned}$$

The rest of the proof is the same as in the proof of Theorem 1. □

5 Barriers for Improving the Approximation Ratio

In this section, we show some barriers for further improving the approximation ratio, even on the special case of Correlation Clustering. We show that there exists an instance of Correlation Clustering and an LP solution such that the algorithm has a cost that is strictly greater than 6 times the LP cost, under the worst merge order. This does not eliminate the possibility that we can use a different merge order or some primal-dual argument to get a better approximation ratio, but it shows that the current simple analysis can not be improved by much.

Let $C(\delta)$ be the supremum over all instances and feasible LP solutions x of the ℓ_1 error of the solution produced by the $\mathsf{Merge}(\delta)$) algorithm divided by $\mathsf{lp\text{-}cost}_p(x)$. We construct instances of Correlation Clustering where each pair $i, j \in [n]$ is labeled either $+$ or $-$, with the intent

that pairs labeled + ought to be in the same cluster and pairs labeled − ought to be in different clusters. For convenience, we use a slightly different terminology in describing the constructions below: every pair has a *target*; the target for a + edge is 0 and the target for a − edge is 1. A solution is a disjoint partition of $[n]$ into clusters and the cost is the number of pairs $i, j \in [n]$ where the clustering deviates from the label for that pair.

5.1 Weighted points and clones

Our constructions are most naturally described with points that carry *masses*, so that a "point" stands for many vertices in the same place. A *weighted instance* consists of a finite set P of points, rational masses $w_p > 0$, a $[0,1]$-semimetric x on P, and a target $z_{pq} \in \{0,1\}$ for every pair of distinct points. For disjoint sets $A, C \subseteq P$ we write $w(A) = \sum_{p \in A} w_p$ and use the weighted average

$$\bar{x}(A,C) = \frac{1}{w(A)w(C)} \sum_{p \in A} \sum_{q \in C} w_p w_q x_{pq}.$$

Weighted legal merge histories for the algorithm are defined with this average. For a partition $\mathcal{P}$ of P, define the weighted mistakes and cost

$$\mathrm{mist}_w(\mathcal{P}) = \sum_{\{p,q\}} w_p w_q \left| y_{pq} - z_{pq} \right|, \qquad \mathsf{cost}_w(x) = \sum_{\{p,q\}} w_p w_q \left| x_{pq} - z_{pq} \right|.$$

Lemma 15 (Clones)**.** *Let (P, w, x, z) be a weighted instance and let $\mathcal{P}$ be the output of a legal weighted history of $\mathsf{Merge}(\delta)$. If $\mathsf{cost}_w(x) > 0$, then $C(\delta) \geq \mathrm{mist}_w(\mathcal{P})/\mathsf{cost}_w(x)$.*

Proof. Choose a positive integer Q such that every Qw_p is an integer. Replace each point p by a set V_p of Qw_p vertices, called its *clones*. Two clones of the same point get fractional distance 0 and target 0; a clone of p and a clone of $q \neq p$ get distance x_{pq} and target z_{pq}. This is a $[0,1]$-semimetric.

For sets A, C of points, write $\widehat{A} = \bigcup_{p \in A} V_p$. Then $\left|\widehat{A}\right| = Qw(A)$ and $\sum_{u \in \widehat{A},\, v \in \widehat{C}} x_{uv} = Q^2 \sum_{p \in A,\, q \in C} w_p w_q x_{pq}$, so the unweighted average distance between $\widehat{A}$ and $\widehat{C}$ equals the weighted average between A and C. Now run $\mathsf{Merge}(\delta)$ as follows. First merge the clones of each point into one cluster V_p; each of these merges has average distance $0 \leq \delta$. Then perform the weighted history, replacing each cluster A by $\widehat{A}$. Every merge is eligible and the stopping condition holds, because all averages agree with the weighted ones.

Pairs of clones of the same point contribute nothing to both the fractional cost and the number of mistakes. Every pair of distinct points p, q corresponds to $Q^2 w_p w_q$ pairs of vertices with the same distance, target, and output. Hence the number of mistakes is $Q^2 \mathrm{mist}_w(\mathcal{P})$ and the fractional cost is $Q^2 \mathsf{cost}_w(x)$. Thus the lemma holds. □

5.2 The main theorem

Throughout, let

$$\mathsf{B} = \frac{2375608}{390625} = 6.08155648, \qquad \delta_- = \frac{1}{\mathsf{B}} = \frac{390625}{2375608}, \qquad \delta_+ = \frac{1 - 1/\mathsf{B}}{2} = \frac{1984983}{4751216},$$

so $\delta_- \approx 0.16443$ and $\delta_+ \approx 0.41778$. Let $\mathcal{D} = [\delta_-, \delta_+]$ be the *middle range*.

Theorem 16. *For every $\delta > 0$ we have $C(\delta) \geq \mathsf{B}$.*

The proof combines three results:

- for $\delta \le \delta_-$: $\Gamma(\delta) \ge 1/\delta \ge \mathsf{B}$ (Lemma 17);
- for $\delta_+ \le \delta < 1/2$: $\Gamma(\delta) \ge 1/(1-2\delta) \ge \mathsf{B}$, and for $\delta \ge 1/2$: $\Gamma(\delta) = \infty$ (Lemma 18);
- for $\delta \in \mathcal{D}$: $\Gamma(\delta) \ge R(\delta) > \mathsf{B}$ (Proposition 31).

5.3 Warm Up: Two easy bounds

We begin with two easy lower bounds.

Lemma 17.

$$C(\delta) \ge \frac{1}{\delta}$$

Proof. Consider an instance with two vertices i, j labeled $+$ with $x_{ij} = \delta + \varepsilon$. The algorithm keeps them separate, with a cost of 1, while the LP cost is $\delta + \varepsilon$. The ratio tends to $\frac{1}{\delta}$ as $\varepsilon \to 0$. □

Lemma 18. *For* $0 < \delta < 1/2$

$$C(\delta) \ge \frac{1}{1-2\delta}$$

Proof. Fix $0 < \theta < \delta$, integers $K, S \ge 1$, and let $a = \theta\frac{K+1}{K+2} < \theta$. The LP solution x_{ij} is obtained from distances on the real line: Place K seed vertices at position 0, S vertices at $-a$ and S vertices at $+a$. Pairs i, j with one vertex at a, and the other at $-a$ are labeled $-$; all other pairs are labeled $+$.

A valid merge order for $\mathsf{Merge}(\delta)$ is as follows: First merge all the seed vertices into one cluster (distances are 0). Then alternately attach one vertex from $-a$ and one from $+a$. Suppose the current cluster contains the seeds and t vertices on each side. Attaching a vertex at $-a$ has average distance

$$\frac{K \cdot a + t \cdot 2a + t \cdot 0}{K + 2t} = a < \theta < \delta.$$

Then attaching a vertex at $-a$ has average distance

$$\frac{K \cdot a + (t+1) \cdot 2a + t \cdot 0}{K + 2t} = a\left(1 + \frac{1}{K + 2t + 1}\right) \le \frac{K+2}{K+1} = \theta < \delta$$

Thus every merge is eligible and the algorithm ends with a single cluster.

The cost of the algorithm is S^2 (the number of pairs labeled $-$). The LP cost of the solution is $S^2(1-2a) + 2KSa$ The ratio is

$$\frac{S^2}{S^2(1-2a) + 2KSa} = \frac{1}{1 - 2a + 2aK/S}$$

Choosing parameters so that $S/K \to \infty$ gives ratios approaching $1/(1-2a)$. We let $K \to \infty$ and $\theta \to \delta$, so $a \to \delta$. Thus the ratio approaches $1/(1-2\delta)$ □

We note that this argument also shows that the ratio is unbounded for $\delta \ge 1/2$.

5.4 Amplification on a star

We build instances from *arms*. An arm is a distribution of points along a segment that starts at a common center, and every point has one of two labels.

Definition 19 (Arm)**.** *An* arm *is a finite list of* atoms (r_k, w_k, λ_k)*, where* $r_k \in [0,1)$ *is a* radius *(distance from the center),* $w_k > 0$ *is a rational mass with* $\sum_k w_k = 1$*, and* $\lambda_k \in \{low, high\}$ *is a* label*. An arm is* buildable at threshold δ *if the weighted instance on its atoms with distances* $|r_k - r_m|$ *has a sequence of merges, ending in a single cluster, in which every merge has weighted average distance at most* δ*.*

It is convenient to view an arm as a probability distribution: choosing atom k with probability w_k gives a random radius r and label λ. Let (r, λ) and (s, λ') be two independent draws from the arm, which we think of as points on two different arms of a star. The key quantities we will need are the *cross mean* μ, the *mistake mass* N, and the *cross cost* F:

$$\mu = \mathsf{E}\min\{r+s, 1\}, \tag{14}$$

$$N = \Pr[\lambda = \lambda' = \text{low}] = \ell^2, \qquad \text{where } \ell = \Pr[\lambda = \text{low}], \tag{15}$$

$$F = \mathsf{E}\big[\kappa(r, \lambda, s, \lambda')\big], \qquad \kappa(r, \lambda, s, \lambda') = \begin{cases} \min\{r+s, 1\} & \text{if } \lambda = \lambda' = \text{low}, \\ 1 - \min\{r+s, 1\} & \text{otherwise.} \end{cases} \tag{16}$$

We write $\beta = 1 - \ell$ for the high mass (for the specific arms below, β will be given by an explicit formula).

Definition 20 (Star instance)**.** *Given an arm and an integer* $M \geq 2$*, the* star instance *has points* (i, k) *for each* arm index $i \in \{1, \ldots, M\}$ *and each atom* k*, with mass* w_k*. Its fractional distances are*

$$x\big((i,k),(j,m)\big) = \begin{cases} |r_k - r_m| & \text{if } i = j, \\ \min\{r_k + r_m, 1\} & \text{if } i \neq j, \end{cases} \tag{17}$$

and its targets are

- $z = 0$ *(positive) for every pair on the same arm;*
- $z = 0$ *(positive) for a pair on different arms if both points are low;*
- $z = 1$ *(negative) for a pair on different arms if at least one point is high.*

Lemma 21. *The distances* (17) *form a* $[0,1]$*-semimetric.*

Proof. Consider the star with a center o and M edges (*rays*) of length one, and place (i, k) on ray i at distance r_k from o. The shortest-path distance is $|r_k - r_m|$ on the same ray and $r_k + r_m$ on different rays. This is a tree metric, so its capping at one is a $[0,1]$-semimetric. On the same ray $|r_k - r_m| < 1$, so the capping does not change these values. □

Points of radius 0 on different arms are at distance 0; they all sit at the center, and this is essential for the construction.

We design the instance so that the final clusters are exactly the arms. For this solution, the mistakes are exactly the cross-arm pairs of two low points which are positive but separated (Figure 2).

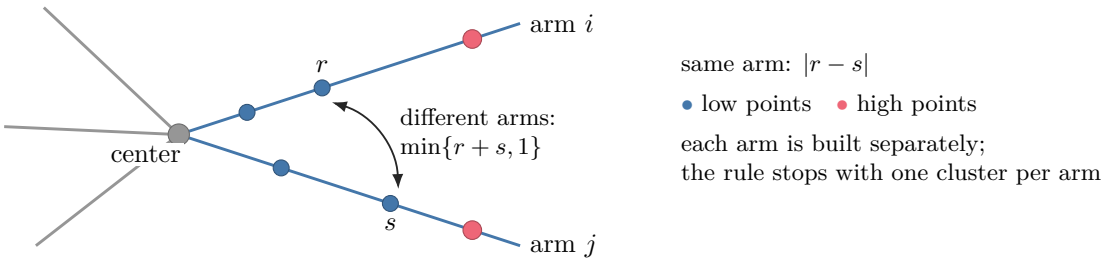


Figure 1: The star instance. Every arm carries an identical copy of the same distribution of points. Fractional distances are path lengths in the star, capped at one. The red high points sit far from the center.

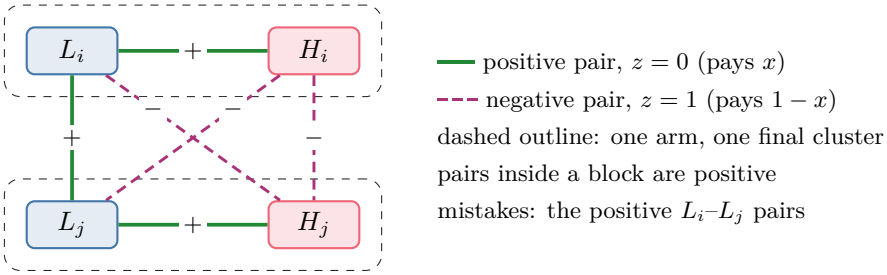


Figure 2: The targets of the star instance, drawn for two arms i, j. Each arm splits into its low part L_i and its high part H_i. The output keeps every arm together and separates different arms, so the mistakes are exactly the low–low pairs between arms.

Lemma 22 (Amplification)**.** *Let $\delta > 0$, and let an arm be buildable at threshold δ with cross mean $\mu > \delta$ and cross cost $F > 0$. Then for every $M \geq 2$,*

$$C(\delta) \;\geq\; \frac{N}{F + \delta/(M-1)}, \qquad \textit{and hence} \qquad C(\delta) \;\geq\; \frac{N}{F}.$$

Proof. Consider the star instance with M arms, and the following sequence of merges: build arm 1 into one cluster using the merges that show that the arm is buildable, then arm 2, and so on. Distances within an arm are the distances $|r_k - r_m|$ of the arm, so every merge has average distance at most δ. Note that other pairs of clusters may be eligible along the way; this does not invalidate our construction of a valid merge history for the algorithm. When all arms are built, the clusters are the M arms. For $i \neq j$, the weighted average distance between arm i and arm j is exactly $\mu > \delta$ by (14), so no pair is eligible and the constructed merge history is indeed valid.

By the discussion before the lemma, the weighted mistakes are the low–low pairs between arms: $\text{mist}_w = \binom{M}{2}\ell^2 = \binom{M}{2}N$. The fractional cost of a pair between arms is κ from (16): positive pairs pay x and negative pairs pay $1-x$. Summing over the $\binom{M}{2}$ pairs of arms gives $\binom{M}{2}F$. It remains to bound the *internal cost* I of one arm,

$$I = \sum_{\{k,m\}} w_k w_m \, |r_k - r_m| \,,$$

the sum over pairs of distinct atoms; these pairs are positive and pay their distance. Assign each pair $\{k, m\}$ to the unique merge of the arm's history that first puts k and m into the same cluster, say a merge of clusters A and C with $k \in A$ and $m \in C$. The pairs assigned to this merge are

exactly those in $A \times C$, and their total contribution is $w(A)w(C)\,\bar{x}(A,C) \le \delta\, w(A)w(C)$. Summing over all merges,

$$I \;\le\; \delta \sum_{\text{merges}} w(A)w(C) = \delta \sum_{\{k,m\}} w_k w_m \;\le\; \frac{\delta}{2}\Big(\sum_k w_k\Big)^2 = \frac{\delta}{2}.$$

So $\mathsf{cost}_w \le \binom{M}{2}F + M\delta/2$, and by Lemma 15,

$$C(\delta) \;\ge\; \frac{\binom{M}{2}N}{\binom{M}{2}F + M\delta/2} = \frac{N}{F + \delta/(M-1)}.$$

Letting $M \to \infty$ gives the second bound. □

Obtaining a large ratio. Lemma 22 reduces the problem to designing a single arm. We want the mistake mass $N = \ell^2$ to be large and the cross cost F to be small, subject to two constraints: the arm must be buildable by eligible merges, and two arms must not be eligible, $\mu > \delta$. Low points near the center make F small (a low–low pair on two arms pays only $r + s$), but they also make μ small. Suppose every high point has radius at least $\frac{1}{2}$. Then two high points on different arms are at distance 1; such a pair is negative, so it costs nothing, and it raises the cross mean μ. The lowhigh pairs on different arms are where the fractional solution pays for using high points to keep the arms apart, even though the algorithm handles those pairs correctly. The next section chooses the distribution of radii that balances these effects.

5.5 The idealized arm and its costs

In this section we describe an idealized arm, given by a distribution with a density, and compute its cross mean and costs exactly. Subsection 5.6 then approximates it by finite buildable arms.

5.5.1 The profile

Fix $\delta \in (0, 1/2)$ and define

$$b = \tfrac{1}{2} - \delta, \qquad H = \tfrac{1}{2} + \delta, \qquad \zeta = e^{-b/\delta} = e^{1-1/(2\delta)}, \qquad \alpha = \frac{1}{1 + \sqrt{1 - \zeta^2/2}}, \qquad \beta = \frac{\alpha\zeta}{2}. \tag{18}$$

Note that $0 < \zeta < 1$, and hence $\frac{1}{2} < \alpha < \frac{1}{1+1/\sqrt{2}} < 0.59$ and $\beta < \frac{1}{2}$.

Definition 23 (Core and profile)**.** *The* core *is the random variable $Y \in [0, H]$ with density $\delta^{-1}e^{-y/\delta}$ on $(0, b)$ and two atoms, of mass $\zeta/2$ each, at b and at H. The* profile *ν_δ is the distribution of the radius r that equals 0 with probability $1 - \alpha$ and is distributed as Y with probability α. The atom of ν_δ at H, of mass $\alpha \cdot \zeta/2 = \beta$, is labelled* high*; everything else is* low*.*

The core has total mass $\int_0^b \delta^{-1}e^{-y/\delta}\,\mathrm{d}y + \zeta = (1 - \zeta) + \zeta = 1$. The two atoms of the core sit symmetrically at distance δ from the point $1/2$ (Figure 3). This is exactly the kind of cluster that the seed construction of Lemma 18 builds: a seed at $1/2$ grows two equal groups at $1/2 \pm \delta$.

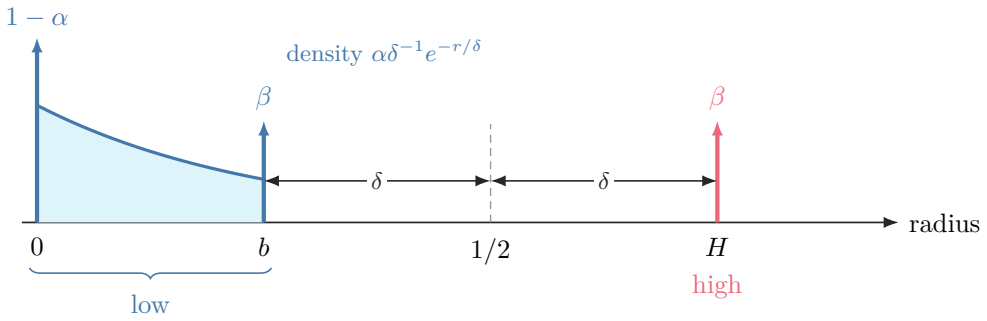


Figure 3: The profile ν_δ, drawn schematically for $\delta = 1/4$. Arrows are point masses; the shaded curve is the density. The two terminal atoms at b and H are symmetric about $1/2$.

5.5.2 Three identities

The exponential density is chosen to make the following *tail identity* hold. In words: if we look only at the part of the core at radius at least t, its average distance from t is always exactly δ.

Lemma 24 (Tail identity)**.** *For every $0 \le t \le b$,*

$$\Pr[Y \ge t] = e^{-t/\delta} \qquad \textit{and} \qquad \mathsf{E}\big[(Y-t)\mathbb{1}\{Y \ge t\}\big] = \delta e^{-t/\delta}.$$

In particular $\mathsf{E}[Y] = \delta$, and therefore $\mathsf{E}_{\nu_\delta}[r] = \alpha\delta$.

Proof. For the first identity, the event $Y \ge t$ consists of the density part on $[t, b)$ and both atoms, so $\Pr[Y \ge t] = (e^{-t/\delta} - e^{-b/\delta}) + \zeta = e^{-t/\delta}$.

For the second, we use $\mathsf{E}[(Y-t)\mathbb{1}\{Y \ge t\}] = \int_t^\infty \Pr[Y > u]\,du$. Now consider various intervals for u. For $t \le u < b$ we have $\Pr[Y > u] = e^{-u/\delta}$ (by the same computation, the point u having no mass). For $b \le u < H$ only the top atom remains, and $\Pr[Y > u] = 0$ for $u \ge H$. Hence the integral is

$$\int_t^b e^{-u/\delta}\,du + (H-b)\frac{\zeta}{2} = \delta\big(e^{-t/\delta} - e^{-b/\delta}\big) + 2\delta \cdot \frac{\zeta}{2} = \delta\big(e^{-t/\delta} - \zeta\big) + 2\delta \cdot \frac{\zeta}{2} = \delta e^{-t/\delta}.$$

Taking $t = 0$ gives $\mathsf{E}[Y] = \delta$. Recall that ν_δ is the distribution of the radius r that equals 0 with probability $1-\alpha$ and is distributed as Y with probability α. Hence, $\mathsf{E}_{\nu_\delta}[r] = \alpha\mathsf{E}[Y] = \alpha\delta$. □

The value of α is chosen so that two arms have cross mean exactly δ, which is established by the following identity.

Lemma 25. *$\alpha^2\zeta^2 = 4\alpha - 2$; equivalently, $\beta^2 = \alpha - \frac{1}{2}$.*

Proof. Put $\sigma = \sqrt{1-\zeta^2/2}$, so $1/\alpha = 1+\sigma$. Squaring $\sigma = 1/\alpha - 1$ gives $1 - \zeta^2/2 = 1/\alpha^2 - 2/\alpha + 1$. Multiply by α^2 and rearrange. The second form follows from $\beta = \alpha\zeta/2$. □

Lemma 26 (Cross mean)**.** *For independent $r, s \sim \nu_\delta$, $\mathsf{E}[\min(r+s, 1)] = \delta$.*

Proof. Low radii are at most b and the high radius is H. Since $2b = 1 - 2\delta < 1$ and $b + H = 1$, the sum $r + s$ exceeds one only when both radii equal H; then $\min(r+s, 1) = r + s - 2\delta$, because $2H - 1 = 2\delta$. Therefore

$$\mathsf{E}[\min(r+s,1)] = \mathsf{E}[r+s] - 2\delta\beta^2 = 2\mathsf{E}_{\nu_\delta}[r] - 2\delta\beta^2 = 2\alpha\delta - 2\delta\Big(\alpha - \frac{1}{2}\Big) = \delta,$$

using Lemmas 24 and 25. □

5.5.3 The costs

We now compute the mistake mass and the cross cost of the profile. Write ν_L for the restriction of ν_δ to its low part (a measure of total mass $\ell = 1 - \beta$) and

$$m_L = \int r \, \mathrm{d}\nu_L(r) = \alpha\delta - \beta H$$

for its first moment; the formula follows from Lemma 24, since the high atom contributes βH to $\mathsf{E}r = \alpha\delta$.

Lemma 27 (Costs of the profile)**.** *For two independent draws from ν_δ, the mistake mass and the cross cost are*

$$N(\delta) = (1-\beta)^2, \tag{19}$$
$$F(\delta) = 2\ell m_L + 2\beta\,(b\ell - m_L) \tag{20}$$
$$= 2\alpha\delta - 4\delta(\alpha+1)\beta + (1+6\delta)\beta^2. \tag{21}$$

Moreover $F(\delta) > 0$.

Proof. Equation (19) is the definition (15). For the cross cost we sort the pairs into three kinds (Figure 4).

- *Low–low.* Since $r+s \le 2b < 1$, the pair pays $r+s$. Its contribution is $\iint (r+s)\,\mathrm{d}\nu_L(r)\,\mathrm{d}\nu_L(s) = 2\ell m_L$.
- *Low–high*, in either order. The distance is $r + H \le b + H = 1$, so the pair pays $1 - (r+H) = b - r \ge 0$. The two orders contribute $2\beta\int (b-r)\,\mathrm{d}\nu_L(r) = 2\beta(b\ell - m_L)$.
- *High–high.* The capped distance is 1 and the target is 1: cost 0.

This proves (20). For (21), substitute $\ell = 1-\beta$ and $m_L = \alpha\delta - \beta(\frac{1}{2} + \delta)$ and expand: we get $2m_L(1-2\beta) + 2\beta b(1-\beta)$, and collecting terms gives (21). Finally $m_L > 0$ because the density part of ν_L has positive mass at positive radii, and $b\ell - m_L = \int (b-r)\,\mathrm{d}\nu_L \ge 0$; hence $F \ge 2\ell m_L > 0$. □

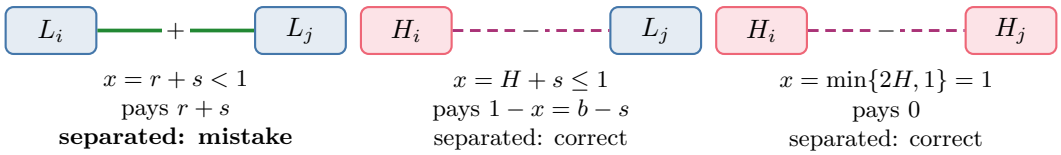


Figure 4: Bookkeeping for a pair of arms. Only low–low pairs are mistakes. The fractional solution pays a little for them (low radii are small) and a little for the mixed pairs, and nothing for high–high pairs.

Let $R(\delta) = N(\delta)/F(\delta)$. Substituting $\beta = \alpha\zeta/2$ and $\beta^2 = \alpha - \frac{1}{2}$ (Lemma 25) into (19) and (21) removes all powers of β:

$$R(\delta) = \frac{N(\delta)}{F(\delta)} = \frac{\alpha + \frac{1}{2} - \alpha\zeta}{\delta\,(8\alpha - 3 - 2\alpha\zeta - 2\alpha^2\zeta) + \alpha - \frac{1}{2}}. \tag{22}$$

For example, at $\delta = 1/4$ we have $b = 1/4$, $H = 3/4$, $\zeta = e^{-1}$, and Lemma 25 becomes $\beta^2 - 2e\beta + \frac{1}{2} = 0$, so $\beta = e - \sqrt{e^2 - 1/2} \approx 0.0935807$ and $\alpha = 2e\beta \approx 0.508757$. Then $N \approx 0.821596$, $F \approx 0.135081$, and $R(1/4) \approx 6.082225$.

Intuition for this probability distribution. The tail identity says that every "top part" of the core has average distance exactly δ from the radius where it starts. In the next section, we will show how this allows the core to be assembled from the top down, starting with the two terminal atoms at b and H, and then, performing a sequence of merges sweeping towards the origin. By design, every merge of the sweep is *saturated*, i.e. has average distance exactly equal to the allowed maximum (Subsection 5.6). A saturated arm is as spread out as possible, which keeps the low radii small relative to the cross mean. Mass at the center (the $1-\alpha$ part) can always be added in a single final merge, since a cluster at the center has average distance $\mathsf{E}Y = \delta$ to the core. It is the cheapest way to create low–low mistakes, because a low–low pair of two center points has fractional distance 0. By Lemma 26, $1-\alpha$ is exactly the amount of dilution at which the cross mean drops to δ; in the actual construction of the finite arms of Subsection 5.6, we dilute slightly less, so that each arm is its own cluster and two such clusters do not merge further.

5.6 Finite arms construction

In this section we build exact finite arms that approximate the idealized construction of the previous section, are buildable (i.e. have a valid merge history), and have $\mu > \delta$. Three small changes make this possible. The threshold δ (which may be irrational) is replaced by a rational parameter $\theta < \delta$, so that all masses are rational and every merge has average distance at most θ. The two terminal groups are placed slightly closer than $\pm\theta$ to the point $\frac{1}{2}$, so that the seed construction of Lemma 18 applies. Finally, the gap $\mu - \delta > 0$ is created by slightly increasing α.

5.6.1 The comb arm

Definition 28 (Comb arm)**.** *Let $\theta, \alpha \in \mathbb{Q}$ with $0 < \theta < 1/2$ and $0 < \alpha < 1$, and let $n, K, J \geq 1$ be integers with $n > (\frac{1}{2} - \theta)/\theta$. Put*

$$b' = \tfrac{1}{2} - \theta, \qquad h = \frac{b'}{n}, \qquad \rho = 1 - \frac{h}{\theta} \in (0,1), \qquad a = \theta\,\frac{K+1}{K+2}, \qquad \tau = \frac{\rho^n}{K+2J}.$$

The core *consists of the following atoms (masses are "core masses"):*

- *the* terminal*: a* seed *at radius $\frac{1}{2}$ with mass $K\tau$; J* lower chunks *at radius $\frac{1}{2} - a$ and J* upper chunks *at radius $\frac{1}{2} + a$, each of mass τ;*
- *the* comb*: for $j = 1, \ldots, n$, an atom at radius jh with mass $(1-\rho)\rho^{j-1}$.*

The comb arm *consists of the core atoms with all masses multiplied by α, together with a* center atom *at radius 0 with mass $1-\alpha$. The upper chunks are labelled high; all other atoms are labelled low.*

All radii and masses are rational, all radii lie in $[0,1)$, and the core has total mass $\rho^n + \sum_{j=1}^{n}(1-\rho)\rho^{j-1} = \rho^n + (1-\rho^n) = 1$, so the arm has total mass one. Chunks at the same radius are distinct atoms at distance 0 from each other; they are the units that are attached one at a time.

Lemma 29 (The comb arm is buildable)**.** *The comb arm has a merge history, ending in a single cluster, in which every merge has weighted average distance at most θ. In particular, it is buildable at every threshold $\delta > \theta$.*

Proof. Multiplying all core masses by α does not change weighted averages between sets of core atoms, so we work with core masses. The merge history has three phases (Figure 5).

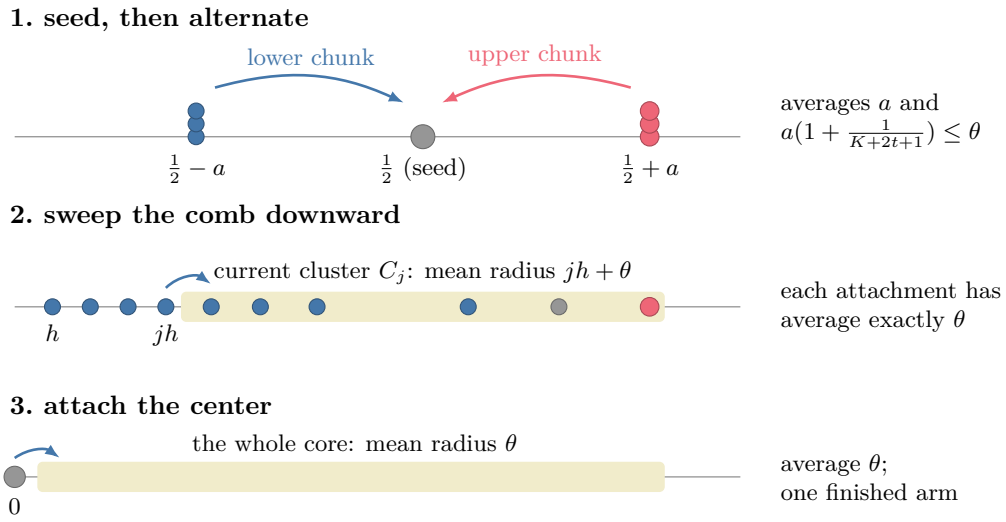


Figure 5: The seed-and-sweep history of a comb arm. Row 1: starting from the seed at $\frac{1}{2}$, lower (blue) and upper (red, high) chunks are attached alternately. Row 2: the comb atoms are attached in order of decreasing radius; the shaded region is the cluster built so far. Row 3: the center atom joins last. Every merge has average distance at most $\theta < \delta$.

Phase 1: the terminal. (Similar to the proof of Lemma 18). Start with the seed. Alternately attach one lower chunk and one upper chunk. If the cluster consists of the seed and t chunks of each kind, the lower chunk is at distance a from the seed, 0 from the lower chunks, and $2a$ from the upper chunks, so its average distance is

$$\frac{K\tau \cdot a + t\tau \cdot 2a + t\tau \cdot 0}{(K+2t)\tau} = a \le \theta.$$

Then the upper chunk has average distance

$$\frac{K\tau \cdot a + (t+1)\tau \cdot 2a + t\tau \cdot 0}{(K+2t+1)\tau} = a\Big(1 + \frac{1}{K+2t+1}\Big) \le a\,\frac{K+2}{K+1} = \theta.$$

After J rounds the cluster is the whole terminal. It has mass ρ^n, all elements have radii (distance from the origin) at least $\frac{1}{2} - a \ge \frac{1}{2} - \theta = nh$, and by symmetry its mean radius is $\frac{1}{2} = nh + \theta$.

Phase 2: the comb. For $j = n, n-1, \ldots, 1$ we attach the comb atom at jh. Let C_j denote the cluster just before this attachment: the terminal together with the comb atoms at ih for $j < i \le n$. We claim that

$$w(C_j) = \rho^j, \qquad \text{mean radius of } C_j = jh + \theta, \qquad \text{all radii in } C_j \text{ are} \ge jh. \tag{23}$$

For $j = n$ this was established at the end of Phase 1. If (23) holds for j, then the atom at jh is at distance (radius $- jh$) from every atom of C_j, so its average distance to C_j is $(jh + \theta) - jh = \theta$. After the attachment, C_{j-1} has mass $\rho^j + (1-\rho)\rho^{j-1} = \rho^{j-1}$, its radii are at least $jh \ge (j-1)h$, and its mean radius is

$$\frac{\rho^j(jh+\theta) + (1-\rho)\rho^{j-1} \cdot jh}{\rho^{j-1}} = jh + \rho\theta = jh + \theta - h = (j-1)h + \theta,$$

using $\rho\theta = \theta - h$. This proves (23) by induction. In particular, C_0 is the whole core and has mean radius θ.

Phase 3: the center. The center atom is at distance (radius) from every core atom, so its average distance to the core is the mean core radius, θ. □

Note that the invariant established in Phase 2 is a discrete version of the tail identity (Lemma 24): as $n \to \infty$, the geometric masses $(1-\rho)\rho^{j-1}$ at jh approach the exponential density $\delta^{-1}e^{-r/\delta}$ when $\theta \approx \delta$.

Lemma 30 (Cross quantities of the comb arm)**.** *For the comb arm, write $\beta = \alpha J\tau$ (the high mass), $\ell = 1-\beta$, $s_0 = \alpha K\tau$ (the seed mass), and $m_L = \alpha\theta - \beta(\frac{1}{2}+a)$. Then*

$$\begin{aligned} \textit{mean } \mu &= 2\alpha\theta - 2a\beta(\beta + s_0), \\ \textit{mistake mass } N &= \ell^2, \\ \textit{cross mean } F &= 2\ell m_L + 2\beta\big((\tfrac{1}{2}-a)\ell - m_L + as_0\big). \end{aligned}$$

Proof. The mean radius of the arm is $\alpha\theta$ (Phase 2 above), and the high atoms have total mass β at radius $\frac{1}{2}+a$, so $m_L = \alpha\theta - \beta(\frac{1}{2}+a)$ is the first moment of the low part. The low radii are 0, the comb radii (at most $nh = \frac{1}{2} - \theta < \frac{1}{2} - a$), $\frac{1}{2}-a$, and the seed radius $\frac{1}{2}$.

Cross mean. We have $\mathsf{E}[r+s] = 2\alpha\theta$. A sum $r+s$ exceeds one only for a high–high pair (sum $1+2a$, excess $2a$, probability β^2) or a seed–high pair in either order (sum $1+a$, excess a, probability $2s_0\beta$). Two seeds give exactly 1 and a lower chunk with a high atom gives exactly 1. Subtracting the excesses gives the formula for μ.

Cross cost. Low–low sums are at most 1, so these pairs contribute $2\ell m_L$. A low–high pair pays $1 - \min\{r + \frac{1}{2} + a, 1\} = (\frac{1}{2} - a - r)^+$, which equals $\frac{1}{2} - a - r$ for every low radius except the seed, where it is 0 instead of $-a$. Hence the two orders contribute $2\beta\big((\frac{1}{2}-a)\ell - m_L + as_0\big)$. High–high pairs have capped distance 1 and cost 0. □

5.7 From finite arms to the profile

Proposition 31. *For every $\delta \in (0, 1/2)$, $C(\delta) \geq R(\delta)$.*

Proof. Let b, H, ζ, α_* denote the quantities of (18) at δ (we write α_* for α there, because α will now be a free variable). For $k \geq 1$ choose a rational $\theta_k \in (\delta - 1/k, \delta)$ with $\theta_k > 0$, an integer $n_k \geq k$ with $n_k \geq 2(\frac{1}{2} - \theta_k)/\theta_k$, and put $K_k = k$, $J_k = k^2$. Let ρ_k, a_k, τ_k be the resulting quantities of Definition 28, and set

$$p_k = J_k\tau_k = \rho_k^{n_k}\frac{k^2}{k+2k^2}, \qquad q_k = K_k\tau_k = \rho_k^{n_k}\frac{k}{k+2k^2}.$$

Limits of the parameters. Put $y_k = (\frac{1}{2} - \theta_k)/\theta_k$, so $\rho_k = 1 - y_k/n_k$ and $y_k \to b/\delta$. Since $y_k/n_k \leq \frac{1}{2}$ and $-u - u^2 \leq \ln(1-u) \leq -u$ for $0 \leq u \leq \frac{1}{2}$, we get $-y_k - y_k^2/n_k \leq n_k \ln \rho_k \leq -y_k$, hence $\rho_k^{n_k} \to e^{-b/\delta} = \zeta$. Consequently

$$p_k \to \zeta/2, \qquad q_k \to 0, \qquad a_k \to \delta, \qquad \theta_k \to \delta.$$

The quantities as functions of α. By Lemma 30, for the comb arm with parameter α the cross mean $\mu_k(\alpha)$, mistake mass $N_k(\alpha)$, and cross cost $F_k(\alpha)$ are polynomials in α of degree at most two (recall $\beta = \alpha p_k$ and $s_0 = \alpha q_k$), whose coefficients are continuous functions of $(\theta_k, a_k, p_k, q_k)$. Hence they converge, uniformly for $\alpha \in [0,1]$, to the polynomials obtained by substituting the limits:

$$\begin{aligned} \mu_\infty(\alpha) &= 2\alpha\delta - \tfrac{1}{2}\delta\alpha^2\zeta^2, \\ N_\infty(\alpha) &= (1 - \tfrac{1}{2}\alpha\zeta)^2, \\ F_\infty(\alpha) &= 2\ell m + 2\beta\big(b\ell - m\big), \quad \text{where } \beta = \tfrac{1}{2}\alpha\zeta,\ \ell = 1-\beta,\ m = \alpha\delta - \beta H. \end{aligned}$$

At $\alpha = \alpha_*$ these are exactly the quantities of the profile: $\mu_\infty(\alpha_*) = \delta$ by Lemma 25 (or Lemma 26), and $N_\infty(\alpha_*)/F_\infty(\alpha_*) = R(\delta)$ with $F_\infty(\alpha_*) > 0$ by Lemma 27. Moreover μ_∞ is strictly increasing on $[0,1]$, since $\mu'_\infty(\alpha) = \delta(2 - \alpha\zeta^2) \geq \delta$.

Choosing α. Fix $\varepsilon > 0$. Since $\alpha_* < 0.59$ and N_∞, F_∞ are continuous, there is $\gamma \in (0, 0.4)$ such that $F_\infty(\alpha) \geq \frac{1}{2}F_\infty(\alpha_*) > 0$ and $N_\infty(\alpha)/F_\infty(\alpha) \geq R(\delta) - \varepsilon$ for all $\alpha \in I := [\alpha_* + \gamma/2, \alpha_* + \gamma]$. By monotonicity, $\mu_\infty(\alpha) \geq \delta + \eta$ on I, where $\eta = \mu_\infty(\alpha_* + \gamma/2) - \delta > 0$. By uniform convergence, for all large k we have, for every $\alpha \in I$, $\mu_k(\alpha) > \delta$, $F_k(\alpha) > 0$, and $N_k(\alpha)/F_k(\alpha) \geq R(\delta) - 2\varepsilon$. Fix such a k and a rational $\alpha \in I$ (note $I \subset (0,1)$).

Conclusion. The comb arm with these parameters is buildable at threshold δ (Lemma 29, as $\theta_k < \delta$), has cross mean $\mu_k(\alpha) > \delta$, and has $F_k(\alpha) > 0$. By Lemma 22, $\Gamma(\delta) \geq N_k(\alpha)/F_k(\alpha) \geq R(\delta) - 2\varepsilon$. Since ε was arbitrary, $\Gamma(\delta) \geq R(\delta)$. □

5.8 The scalar inequality

By Proposition 31, it remains to show $R(\delta) > \mathsf{B}$ on the middle range $\mathcal{D}$. Figure 6 shows that this is true with a margin that is comfortable at the ends of $\mathcal{D}$ but tiny near its minimum: R has a minimum of about 6.0815564845 near $\delta \approx 0.24745$, only about $4.5 \cdot 10^{-9}$ above B. (The constant B was chosen as a short rational just below this minimum.)

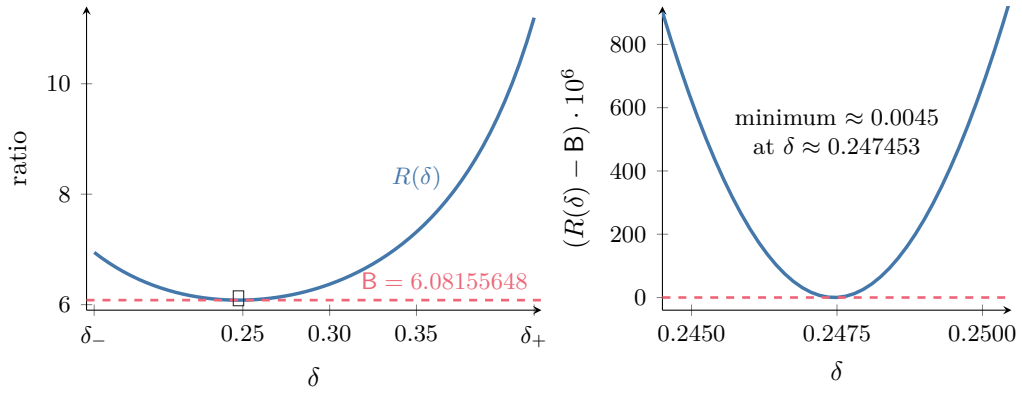


Figure 6: Left: the ratio $R(\delta)$ of the star construction on the middle range (solid), compared with B (dashed). Right: a zoom into the small box, showing the difference $R(\delta) - \mathsf{B}$ magnified by 10^6.

5.9 Clearing the denominator

Since $F(\delta) > 0$, the inequality $R(\delta) > \mathsf{B}$ is equivalent to $N(\delta) - \mathsf{B}F(\delta) > 0$. Define, with ζ and α as in (18),

$$G(\delta) = \delta\big(3 - 8\alpha + 2\alpha\zeta + 2\alpha^2\zeta\big) - \alpha + \frac{1}{2} + \frac{1}{\mathsf{B}}\Big(\frac{1}{2} + \alpha - \alpha\zeta\Big). \tag{24}$$

Comparing with the numerator and denominator of (22),

$$N(\delta) - \mathsf{B}F(\delta) = \mathsf{B}\,G(\delta). \tag{25}$$

So it suffices to show that $G > 0$ on $\mathcal{D}$. Note that G is an explicit function of one real variable.

The inequality is verified by a computer.

5.9.1 Proof of the lower bound

Proof of Theorem 16. Let $\delta > 0$. If $\delta \le \delta_-$, then $C(\delta) \ge 1/\delta \ge \mathsf{B}$ by Lemma 17. If $\delta \ge 1/2$, then $C(\delta) = \infty$ by Lemma 18. If $\delta_+ \le \delta < 1/2$, then $C(\delta) \ge 1/(1-2\delta) \ge \mathsf{B}$ by Lemma 18. Finally, if $\delta \in \mathcal{D}$, then $G(\delta) > 0$, so $N(\delta) > \mathsf{B}F(\delta)$ by (25); since $F(\delta) > 0$ this means $R(\delta) > \mathsf{B}$, and Proposition 31 gives $\mathsf{C}(\delta) \ge R(\delta) > \mathsf{B}$. □

AI Usage

The results in this work were obtained over multiple interactions with GPT-6 Astra. When prompted multiple times to improve the approximation factor for fitting ultrametrics with ℓ_2 error, it initially obtained an $O(\sqrt{\log n})$ approximation, which it subsequently improved to $O(1)$, and extended to all $p \ge 1$. The present algorithm and analysis was obtained over multiple rounds of simplification. The authors suggested a connection to factor revealing LPs for the analysis and this resulted in some further improvements to the approximation factors. The authors used GPT-6 Astra and Claude Opus 5.5 to simplify the analyses as much as possible, then eventually read and understood the results and wrote them up.

References

[ABF+99] Richa Agarwala, Vineet Bafna, Martin Farach, Mike Paterson, and Mikkel Thorup. "On the Approximability of Numerical Taxonomy (Fitting Distances by Tree Metrics)". In: *SIAM Journal on Computing* 28.3 (1999), pp. 1073–1085. DOI: 10.1137/S0097539795296334. URL: https://doi.org/10.1137/S0097539795296334.

[AC11] Nir Ailon and Moses Charikar. "Fitting Tree Metrics: Hierarchical Clustering and Phylogeny". In: *SIAM Journal on Computing* 40.5 (2011). Preliminary version in FOCS 2005, pp. 1275–1291. DOI: 10.1137/100806886. URL: https://doi.org/10.1137/100806886.

[ACN08] Nir Ailon, Moses Charikar, and Alantha Newman. "Aggregating inconsistent information: Ranking and clustering". In: *J. ACM* 55.5 (2008), 23:1–23:27.

[AKLL25] Hyung-Chan An, Mong-Jen Kao, Changyeol Lee, and Mu-Ting Lee. "Handling LP-Rounding for Hierarchical Clustering and Fitting Distances by Ultrametrics". In: *Proceedings of the 66th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2025, pp. 1146–1171. DOI: 10.1109/FOCS63196.2025.00060. URL: https://arxiv.org/abs/2504.06700.

[BBC04] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. "Correlation clustering". In: *Machine learning* 56 (2004), pp. 89–113.

[BDST08] Justin Brickell, Inderjit S Dhillon, Suvrit Sra, and Joel A Tropp. "The metric nearness problem". In: *SIAM Journal on Matrix Analysis and Applications* 30.1 (2008), pp. 375–396.

[CCL+24] Nairen Cao, Vincent Cohen-Addad, Euiwoong Lee, Shi Li, Alantha Newman, and Lukas Vogl. "Understanding the Cluster Linear Program for Correlation Clustering". In: *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC)*. See the corrected full version, arXiv:2404.17509v3. ACM, 2024, pp. 1605–1616. DOI: 10.1145/3618260.3649749. URL: https://arxiv.org/abs/2404.17509v3.

[CDK+24] Vincent Cohen-Addad, Debarati Das, Evangelos Kipouridis, Nikos Parotsidis, and Mikkel Thorup. "Fitting Distances by Tree Metrics Minimizing the Total Error within a Constant Factor". In: *Journal of the ACM* 71.2 (2024). Preliminary version in FOCS 2021, 10:1–10:41. DOI: 10.1145/3639453. URL: https://arxiv.org/abs/2110.02807.

[CDL21] Vincent Cohen-Addad, Rémi De Joannis De Verclos, and Guillaume Lagarde. "Improving ultrametrics embeddings through coresets". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 2060–2068.

[CE67] Luigi L Cavalli-Sforza and Anthony WF Edwards. "Phylogenetic analysis. Models and estimation procedures". In: *American journal of human genetics* 19.3 Pt 1 (1967), p. 233.

[CFLM25] Vincent Cohen-Addad, Chenglin Fan, Euiwoong Lee, and Arnaud de Mesmay. "Fitting Metrics and Ultrametrics with Minimum Disagreements". In: *SIAM Journal on Computing* 54.1 (2025). Preliminary version in FOCS 2022, pp. 92–133. DOI: 10.1137/22M1520190. URL: https://arxiv.org/abs/2208.13920.

[CG24] Moses Charikar and Ruiquan Gao. "Improved Approximations for Ultrametric Violation Distance". In: *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 2024, pp. 1704–1737. DOI: 10.1137/1.9781611977912.68. URL: https://arxiv.org/abs/2311.04533.

[CGW05] Moses Charikar, Venkatesan Guruswami, and Anthony Wirth. "Clustering with qualitative information". In: *J. Comput. Syst. Sci.* 71.3 (2005), pp. 360–383.

[CKL20] Vincent Cohen-Addad, CS Karthik, and Guillaume Lagarde. "On efficient low distortion ultrametric embedding". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 2078–2088.

[CLLN23] Vincent Cohen-Addad, Euiwoong Lee, Shi Li, and Alantha Newman. "Handling Correlated Rounding Error via Preclustering: A 1.73-approximation for Correlation Clustering". In: *Proceedings of the 64th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2023, pp. 1082–1104. DOI: 10.1109/FOCS57990.2023.00065. URL: https://arxiv.org/abs/2309.17243.

[CLN22] Vincent Cohen-Addad, Euiwoong Lee, and Alantha Newman. "Correlation Clustering with Sherali-Adams". In: *63rd IEEE Annual Symposium on Foundations of Computer Science*. 2022, pp. 651–661.

[CMSY15] Shuchi Chawla, Konstantin Makarychev, Tselil Schramm, and Grigory Yaroslavtsev. "Near Optimal LP Rounding Algorithm for CorrelationClustering on Complete and Complete k-partite Graphs". In: *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*. 2015, pp. 219–228.

[Day87] William HE Day. "Computational complexity of inferring phylogenies from dissimilarity matrices". In: *Bulletin of mathematical biology* 49.4 (1987), pp. 461–467.

[DPS15] Marco Di Summa, David Pritchard, and Laura Sanità. "Finding the closest ultrametric". In: *Discrete Applied Mathematics* 180 (2015), pp. 70–80.

[Far72] James S Farris. "Estimating phylogenetic trees from distance matrices". In: *The American Naturalist* 106.951 (1972), pp. 645–668.

[FGR+20] Chenglin Fan, Anna C Gilbert, Benjamin Raichel, Rishi Sonthalia, and Gregory Van Buskirk. "Generalized Metric Repair on Graphs". In: *17th Scandinavian Symposium and Workshops on Algorithm Theory (SWAT 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik. 2020.

[FKW95] Martin Farach, Sampath Kannan, and Tandy Warnow. "A Robust Model for Finding Optimal Evolutionary Trees". In: *Algorithmica* 13.1–2 (1995). Preliminary version in STOC 1993, pp. 155–179. DOI: 10.1007/BF01188585. URL: https://link.springer.com/article/10.1007/BF01188585.

[FRV22] Chenglin Fan, Benjamin Raichel, and Gregory Van Buskirk. "Metric Violation Distance: Hardness and Approximation". In: *Algorithmica* 84.5 (2022). Preliminary version in SODA 2018, pp. 1441–1465. DOI: 10.1007/s00453-022-00940-0.

[GJ17] Anna C Gilbert and Lalit Jain. "If it ain't broke, don't fix it: Sparse metric repair". In: *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2017, pp. 612–619.

[GS26] David García-Soriano and Antoine Schohn. *Approximate Dual Separation for the Cluster LP: a 1.387 approximation for Correlation Clustering*. Version 2. 2026. DOI: 10.48550/arXiv.2607.27829. arXiv: 2607.27829 [cs.DS]. URL: https://arxiv.org/abs/2607.27829v2.

[Har67] John A Hartigan. "Representation of similarity matrices by trees". In: *Journal of the american Statistical Association* 62.320 (1967), pp. 1140–1158.

[HKM05] Boulos Harb, Sampath Kannan, and Andrew McGregor. "Approximating the Best-Fit Tree Under L_p Norms". In: *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM)*. Vol. 3624. Springer, 2005, pp. 123–133. DOI: 10.1007/11538462_11.

[JJS67] CJ Jardine, Nicholas Jardine, and Robin Sibson. "The structure and construction of taxonomic hierarchies". In: *Mathematical Biosciences* 1.2 (1967), pp. 173–179.

[Joh67] Stephen C Johnson. "Hierarchical clustering schemes". In: *Psychometrika* 32.3 (1967), pp. 241–254.

[JS71] Nicholas Jardine and Robin Sibson. *Mathematical taxonomy*. Tech. rep. 1971.

[Kip23] Evangelos Kipouridis. "Fitting Tree Metrics with Minimum Disagreements". In: *31st Annual European Symposium on Algorithms (ESA 2023)*. Vol. 274. 2023, 70:1–70:10. DOI: 10.4230/LIPIcs.ESA.2023.70.

[KM86] Mirko Kivánek and Jaroslav Morávek. "NP-hard problems in hierarchical-tree clustering". In: *Acta informatica* 23 (1986), pp. 311–323.

[Man99] Rosario N Mantegna. "Hierarchical structure in financial markets". In: *The European Physical Journal B-Condensed Matter and Complex Systems* 11 (1999), pp. 193–197.

[RTV86] Rammal Rammal, Gérard Toulouse, and Miguel Angel Virasoro. "Ultrametricity for physicists". In: *Reviews of Modern Physics* 58.3 (1986), p. 765.

[SS62] Peter HA Sneath and Robert R Sokal. "Numerical taxonomy". In: *Nature* 193 (1962), pp. 855–860.

[SWW17] Anastasios Sidiropoulos, Dingkang Wang, and Yusu Wang. "Metric embeddings with outliers". In: *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM. 2017, pp. 670–689.

[WSSB77] M.S. Waterman, T.F Smith, M. Singh, and W.A. Beyer. "Additive evolutionary trees". In: *Journal of Theoretical Biology* 64(2) (1977), pp. 199–213.

A Omitted Proofs in Lemma 13

For the second term on the RHS of Equation (12), we have

$$\begin{aligned}\sum_{ij\in A\times B}(X(i,j)-D_{\ell^*})_+^p &= \sum_{ij\in A\times B}(D_{\mathsf{idx}(i,j)}-D_{\ell^*})_+^p \\ &= \sum_{ij\in A\times B}\sum_{\ell=\ell^*}^{\mathsf{idx}(i,j)-1}((D_{\ell+1}-D_{\ell^*})^p-(D_\ell-D_{\ell^*})^p) \\ &= \sum_{\ell=\ell^*}^{L-1}((D_{\ell+1}-D_{\ell^*})^p-(D_\ell-D_{\ell^*})^p)\cdot|E_-^\ell(A,B)|\ .\end{aligned}$$

Consider any $\ell\geq\ell^*$. Because we have $x_{\ell^*}(A,B)\leq\delta$ and our LP solution ensure that $x_\ell(A,B)$ is decreasing in terms of ℓ (Equation (4)), we have $x_\ell(A,B)\leq\delta$ for any $\ell\geq\ell^*$. Applying Lemma 14, we further have

$$\begin{aligned}&\sum_{ij\in A\times B}(X(i,j)-D_{\ell^*})_+^p \\ &\leq\sum_{\ell=\ell^*}^{L-1}((D_{\ell+1}-D_{\ell^*})^p-(D_\ell-D_{\ell^*})^p)\cdot\Gamma_0\cdot\mathsf{err}_\ell(A,B) \\ &=\Gamma_0\sum_{\ell=\ell^*}^{L-1}((D_{\ell+1}-D_{\ell^*})^p-(D_\ell-D_{\ell^*})^p)\cdot\sum_{ij\in A\times B}\mathbb{E}[|\mathbb{1}(X(i,j)>D_\ell)-\mathbb{1}(Y(i,j)>D_\ell)|] \\ &=\Gamma_0\sum_{ij\in A\times B}\mathbb{E}\left[\sum_{\ell=\ell^*}^{L-1}((D_{\ell+1}-D_{\ell^*})^p-(D_\ell-D_{\ell^*})^p)\cdot|\mathbb{1}(X(i,j)>D_\ell)-\mathbb{1}(Y(i,j)>D_\ell)|\right]\ .\end{aligned}$$

If i,j and $Y(i,j)$ are fixed, then we have a fixed sign for the term $\mathbb{1}(Y(i,j)>D_\ell)-\mathbb{1}(X(i,j)>D_\ell)$ for any $\ell\in[L]$. Therefore, we can move the absolute value outside of the summation over ℓ.

$$\begin{aligned}&\sum_{\ell=\ell^*}^{L-1}((D_{\ell+1}-D_{\ell^*})^p-(D_\ell-D_{\ell^*})^p)\cdot|\mathbb{1}(X(i,j)>D_\ell)-\mathbb{1}(Y(i,j)>D_\ell)| \\ &\quad=\left|\sum_{\ell=\ell^*}^{L-1}((D_{\ell+1}-D_{\ell^*})^p-(D_\ell-D_{\ell^*})^p)\cdot(\mathbb{1}(X(i,j)>D_\ell)-\mathbb{1}(Y(i,j)>D_\ell))\right|\end{aligned}$$

Note that $\sum_{\ell=\ell^*}^{L-1}((D_{\ell+1}-D_{\ell^*})^p-(D_\ell-D_{\ell^*})^p)\,\mathbb{1}(X(i,j)>D_\ell)$ telescopes to $(X(i,j)-D_{\ell^*})_+^p$. Similarly, this holds for the term with $Y(i,j)$. Therefore, we have

$$\sum_{ij\in A\times B}(X(i,j)-D_{\ell^*})_+^p\leq\Gamma_0\sum_{ij\in A\times B}\mathbb{E}\left[\left|(X(i,j)-D_{\ell^*})_+^p-(Y(i,j)-D_{\ell^*})_+^p\right|\right]\ .$$

Note that for any $x, y \in \mathbb{R}_{\geq 0}$, we have $|x^p - y^p| \leq (x + |x-y|)^p - x^p$. Therefore, we have

$$\sum_{ij \in A \times B} (X(i,j) - D_{\ell^*})_+^p \leq \Gamma_0 \sum_{ij \in A \times B} \left(\mathbb{E}\left[((X(i,j) - D_{\ell^*})_+ + |X(i,j) - Y(i,j)|)^p \right] - (X(i,j) - D_{\ell^*})_+^p \right)$$

$$\iff \quad \frac{\Gamma_0 + 1}{\Gamma_0} \sum_{ij \in A \times B} (X(i,j) - D_{\ell^*})_+^p \leq \sum_{ij \in A \times B} \mathbb{E}\left[((X(i,j) - D_{\ell^*})_+ + |X(i,j) - Y(i,j)|)^p \right] \ .$$

Using the Minkowski's inequality, we can upper bound the RHS of the above inequality by

$$\left(\left(\sum_{ij \in A \times B} (X(i,j) - D_{\ell^*})_+^p \right)^{1/p} + \left(\sum_{ij \in A \times B} \mathbb{E}[|X(i,j) - Y(i,j)|^p] \right)^{1/p} \right)^p$$

By take p-th root on both sides, and rearranging the terms, we have

$$\begin{aligned} \left(\left(\frac{\Gamma_0 + 1}{\Gamma_0} \right)^{1/p} - 1 \right) \left(\sum_{ij \in A \times B} (X(i,j) - D_{\ell^*})_+^p \right)^{1/p} &\leq \left(\sum_{ij \in A \times B} \mathbb{E}[|X(i,j) - Y(i,j)|^p] \right)^{1/p} \\ &= (\mathsf{lp}(A \times B, x))^{1/p} \ . \end{aligned}$$

This gives an upper bound of $\frac{\Gamma}{2} \cdot (\mathsf{lp}(A \times B, x))^{1/p}$ for the second term on the RHS of Equation (12).