Maximum gradient embeddings and monotone clustering

Manor Mendel* Assaf Naor†

Abstract

Let (X, d_X) be an n-point metric space. We show that there exists a distribution $\mathscr{D}$ over non-contractive embeddings into trees $f : X \to T$ such that for every $x \in X$,

$$\mathbb{E}_{\mathscr{D}}\left[\max_{y\in X\setminus\{x\}} \frac{d_T(f(x), f(y))}{d_X(x,y)}\right] \le C(\log n)^2,$$

where C is a universal constant. Conversely we show that the above quadratic dependence on $\log n$ cannot be improved in general. Such embeddings, which we call *maximum gradient embeddings*, yield a framework for the design of approximation algorithms for a wide range of clustering problems with monotone costs, including fault-tolerant versions of k-median and facility location.

1 Introduction

Metric embeddings are an invaluable tool in analysis, Riemannian geometry, group theory, graph theory, and the design of approximation algorithms. In most cases embeddings are used to "simplify" a geometric object that we wish to understand, or on which we need to preform certain algorithmic tasks. Thus one tries to "faithfully" represent a metric space as a subset of another space with controlled geometry, whose structure is well enough understood to successfully address the problem at hand. There is some obvious flexibility in this approach: Both the choice of target space and the notion of faithfulness of an embedding can be adapted to the problem we wish to solve. Of course, once these choices are made, the main difficulty is the construction of the required embedding, and in the algorithmic context we have the additional burden of making sure that the embedding can be computed efficiently.

The present paper is inspired by problems from mathematical analysis, but its main motivation is algorithmic. We introduce a new notion of embedding, called *maximum gradient embeddings*, which is "just right" for approximating a wide range of clustering problems. We will then provide optimal maximum gradient embeddings of general finite metric spaces, and use them to design the best known approximation algorithms for several natural clustering problems. We do not attempt to explore here all the possible applications of maximum gradient embeddings, and we suspect that there are many more situations in which our method is applicable. Indeed, rather than being encyclopedic, the main emphasis of the present paper is that these embeddings yield a generic approach to many problems, and we give some example that illustrate this fact. In addition, our work raises interesting algorithmic questions which deserve further investigation.

The flexibility that was described above in the choice of notions of embedding has been exploited to great success by numerous authors in the past four decades. Due to the vast amount of work on this topic, we will

*Computer Science Division, The Open University of Israel, 108 Ravutski St., P.O. Box 808, Raanana 43107, Israel. Email mendelma@gmail.com.

†Theory Group, Microsoft Research, One Microsoft Way, Redmond WA, 98052-6399, USA. Email anaor@microsoft.com.

not give references to the many embedding notions that were used in the mathematical and computer science literature. We wish to stress, however, that apart from the most studied problem of bi-Lipschitz embeddings into Euclidean space, there are a lot of variants of this problem which are useful in many different contexts in mathematics and computer science. These notions include L_1 embeddings, embeddings into dominated L_1 metrics, low dimensional embeddings and dimension reduction, volume preserving embeddings, Fréchet embeddings, snowflake, quasi-isometric, coarse, uniform, and quasisymmetric embeddings, embeddings into hyperbolic spaces, Tits buildings and symmetric spaces, embeddings of subsets, quotients, subsets of quotients and quotients of subsets, embeddings into distributions over trees, ultrametrics and spanning trees, multi-embeddings, embeddings with slack, nearest neighbor preserving embeddings, spanners, additive distortion, and various notions of average distortion. In this paper the target spaces will be distributions over tree metrics (also known as products of trees in the mathematical literature). Our notion of faithfulness is new, and will be described below.

Due to their special structure, it is natural to try to embed metric spaces into trees. This is especially important for algorithmic purposes, as many hard problems are tractable on trees. Unfortunately, this is too much to hope for in the bi-Lipschitz category: As shown by Rabinovich and Raz [38] the n-cycle incurs distortion $\Omega(n)$ in any embedding into a tree. However, one can relax this idea and look for a *random* embedding into a tree which is faithful on average. Such an approach has been developed in recent years by mathematicians and computer scientists. In the mathematical literature this is referred to as embeddings into products of trees, and it is an invaluable tool in the study of negatively curved spaces (see for example [11, 17, 36]).

Probabilistic embeddings into trees became an important algorithmic paradigm due to the work of Bartal [3, 4] (see also [1, 18] for the related problem of embedding graphs into distributions over spanning trees). Bartal's work has led to the design of numerous approximation algorithms for a wide range of NP hard problems. In some cases the best known approximation factors are due to the "probabilistic tree" approach, while in other cases improved algorithms have been subsequently found after the original application of probabilistic embeddings was discovered. But, in both cases it is clear that the strength of Bartal's approach is that it is generic: For a certain type of problem one can quickly get a polylogarithmic approximation using probabilistic embedding into trees, and then proceed to analyze certain particular cases if one desires to find better approximation guarantees. However, probabilistic embeddings into trees do not always work. In [7] Bartal and Mendel introduced the weaker notion of multi-embeddings, and used it to design improved algorithms for special classes of metric spaces. Here we *strengthen* this notion to maximum gradient embeddings, and use it to design approximation algorithms for harder problems to which regular probabilistic embeddings do not apply. From a mathematical viewpoint, analyzing this stronger notion of embedding (which is nevertheless weaker than bi-Lipschitz embedding) creates new challenges, and arguably leads to some surprising results.

Let (X, d_X) and (Y, d_Y) be metric spaces, and fix a mapping $f : X \to Y$. We shall say that f is *non-contractive* if for every $x, y \in X$ we have $d_Y(f(x), f(y)) \geq d_X(x, y)$. The *maximum gradient* of f at a point $x \in X$ is defined as

$$|\nabla f(x)|_\infty = \sup_{y \in X \setminus \{x\}} \frac{d_Y(f(x), f(y))}{d_X(x, y)}. \tag{1}$$

Thus the *Lipschitz constant* of f is given by

$$\|f\|_{\mathrm{Lip}} = \sup_{x \in X} |\nabla f(x)|_\infty.$$

Note that in the mathematical literature, mostly in the context of the study of isoperimetry on general geodesic metric measure spaces (see for example [8, 30]), it is common to define the *modulus of the gradient* of f at $x \in X$ as

$$|\nabla f(x)| = \limsup_{y \to x} \frac{d_Y(f(x), f(y))}{d_X(x, y)}. \tag{2}$$

The definition in (2) is very natural in the context of connected metric spaces, but in the context of finite metric spaces it clearly makes more sense to deal with the maximum gradient as defined in (1).

In what follows when we refer to a tree metric we mean the shortest-path metric on a graph-theoretical tree with weighted edges. Recall that (U, d_U) is an ultrametric if for every $u, v, w \in U$ we have $d_U(u, v) \le \max\{d_U(u, w), d_U(w, v)\}$. It is well known that ultrametrics are tree metrics. The following result is due to Fakcharoenphol, Rao and Talwar [19], and is a slight improvement over an earlier theorem of Bartal [4]. For every n-point metric space (X, d_X) there is a distribution $\mathscr{D}$ over non-contractive embeddings into ultrametrics $f : X \to U$ such that

$$\max_{\substack{x,y \in X \\ x \neq y}} \mathbb{E}_{\mathscr{D}} \frac{d_U(f(x), f(y)}{d_X(x, y)} = O(\log n). \tag{3}$$

The logarithmic upper bound in (3) cannot be improved in general.

Inequality (3) is extremely useful for optimization problems whose objective function is linear in the distances, since by linearity of expectation it reduces such tasks to trees, with only a logarithmic loss in the approximation guarantee. When it comes to non-linear problems, the use of (3) is very limited. We will show that this issue can be addressed using the following theorem, which is our main result.

Theorem 1. *Let (X, d_X) be an n-point metric space. Then there exists a distribution $\mathscr{D}$ over non-contractive embeddings into ultrametrics $f : X \to U$ (thus both the ultrametric (U, d_U) and the mapping f are random) such that for every $x \in X$,*

$$\mathbb{E}_{\mathscr{D}} |\nabla f(x)|_\infty \le C(\log n)^2,$$

where C is a universal constant.

On the other hand there exists a universal constant $c > 0$ and arbitrarily large n-point metric spaces Y_n such that for any distribution over non-contractive embeddings into trees $f : Y_n \to T$ there is necessarily some $x \in Y_n$ for which

$$\mathbb{E}_{\mathscr{D}} |\nabla f(x)|_\infty \ge c(\log n)^2.$$

We call embeddings as in Theorem 1, i.e. embeddings with small expected maximum gradient, *maximum gradient embeddings into distributions over trees* (in what follows we will only deal with distributions over trees, so we will drop the last part of this title when referring to the embedding, without creating any ambiguity). The proof of the upper bound in Theorem 1 is a modification of an argument of Fakcharoenphol, Rao and Talwar [19], which is based on ideas from [3, 12]. It uses the same stochastic decomposition of metric spaces as in [19], but it relies on properties of it which are well known to experts, yet have not been exploited in full strength in previous applications. The argument appears in Section 2. Alternative proofs of the main technical step of the proof of the upper bound in Theorem 1 can be also deduced from the results of [34] or an argument in the proof of Lemma 2.1 in [22]. In both of these references the required inequality is deduced from an analysis of the specific stochastic decomposition of Calinescu, Karloff and Rabani [12] that was used in [19]. Here we present a different approach, which shows that the "padding inequality" proved by Fakcharoenphol, Rao and Talwar in [19] can be used as a "black box" to yield a maximum gradient embedding, and there is no need to recall how the stochastic decomposition was originally defined. Our proof can be viewed as a variant of Talagrand's *generic chaining* method [41] which was originally developed as a tool for bounding similar quantities in the context of Gaussian processes. Indeed, our idea to consider maximum gradient embeddings is partially motivated by Talagrand's work. Moreover, our proof of the upper bound in Theorem 1 is a "decomposition into annuli" argument which is reminiscent of Talagrand's approach.

The heart of this paper is the lower bound in Theorem 1—we found it somewhat surprising, as there are not many natural geometric problems whose answer is such a quadratic dependence on $\log n$. Our proof of this lower bound is much more involved than the proof of the upper bound. The metrics Y_n in Theorem 1 are the diamond graphs of Newman and Rabinovich [37], which will be defined in Section 3. These graphs have been used as counter-examples in several embedding problems— see [10, 23, 31, 37]. In particular, we were inspired to consider these examples by the proof in [23] of the fact that they require distortion $\Omega(\log n)$ in any probabilistic embedding into trees. However, our proof of the $\Omega((\log n)^2)$ lower bound in Theorem 1 is considerably more delicate than the proof in [23]. This proof, together with other lower bounds for maximum gradient embeddings, is presented in Section 3.

1.1 A framework for clustering problems with monotone costs

We now turn to some algorithmic applications of Theorem 1. The general reduction in Theorem 2 below should also be viewed as an explanation why maximum gradient embeddings are so natural— they are precisely the notion of embedding which allows such reductions to go through. As remarked above, we do not attempt to be encyclopedic here— we will present a general paradigm for non-linear clustering problems, and analyze some examples. There are many more directions for future research that arise from the ideas described here. In particular, in a future paper we intend to apply maximum gradient embeddings to problems in online algorithms.

A very general setting of the clustering problem is as follows. Let X be an n-point set, and denote by $\mathrm{MET}(X)$ the set of all metrics on X. A *possible clustering solution* consists of sets of the form $\{(x_1, C_1), \ldots, (x_k, C_k)\}$ where $x_1, \ldots, x_k \in X$ and $C_1, \ldots, C_k \subseteq X$. We think of $C_1, \ldots, C_k$ as the clusters, and x_i as the "center" of C_i. In this general framework we do not require that the clusters cover X, or that they are pairwise disjoint, or that they contain their centers. Thus the space of possible clustering solution is $2^{X\times 2^X}$. Assume that for every point $x \in X$, every metric $d \in \mathrm{MET}(X)$, and every possible clustering solution $P \in 2^{X\times 2^X}$, we are given $\Gamma(x, d, P) \in [0, \infty]$, which we think of as a measure of the dissatisfaction of x with respect to P and d. Our goal is to minimize the average dissatisfaction of the points of X. Formally, given a measure of dissatisfaction (which we also call in what follows a *clustering cost function*) $\Gamma : X \times \mathrm{MET}(X) \times 2^{X\times 2^X} \to [0, \infty]$, we wish to compute for a given metric $d \in \mathrm{MET}(X)$ the value

$$\mathrm{Opt}_\Gamma(X, d) \stackrel{\mathrm{def}}{=} \min\left\{\sum_{x\in X} \Gamma(x, d, P) : P \in 2^{X\times 2^X}\right\}$$

(Since we are mainly concerned with the algorithmic aspect of this problem, we assume from now on that Γ can be computed efficiently.)

We make two natural assumptions on the cost function Γ. First of all, we will assume that it scales homogeneously with respect to the metric, i.e. for every $\lambda > 0$, $x \in X$, $d \in \mathrm{MET}(X)$ and $P \in 2^{X\times 2^X}$ we have $\Gamma(x, \lambda d, P) = \lambda\Gamma(x, d, P)$. Secondly we will assume that Γ is monotone with respecting to the metric, i.e. if $d, \overline{d} \in \mathrm{MET}(X)$ and $x \in X$ satisfy $d(x, y) \le \overline{d}(x, y)$ for every $y \in X$ then $\Gamma(x, d, P) \le \Gamma(x, \overline{d}, P)$. In other words, if all the points in X are further with respect to $\overline{d}$ from x then they are with respect to d, then x is more dissatisfied. This is a very natural assumption to make, as most clustering problems look for clusters which are small in various (metric) senses. We call clustering problems with Γ satisfying these assumptions *monotone clustering problems*. A large part of the clustering problems that have been considered in the literature fall into this framework— we will see some examples presently.

The following theorem is a simple application of Theorem 1. It shows that it is enough to solve monotone clustering problems on ultrametrics, with only a polylogarithmic loss in the approximation factor.

Theorem 2 (reduction to ultrametrics)**.** *Let X be an n-point set and fix a homogeneous monotone clustering cost function $\Gamma : X \times \mathrm{MET}(X) \times 2^{X \times 2^X} \to [0,\infty]$. Assume that there is a randomized polynomial time algorithm which approximates $\mathrm{Opt}_\Gamma(X,\rho)$ to within a factor $\alpha(n)$ on any ultrametric $\rho \in \mathrm{MET}(X)$. Then there is a polynomial time algorithm which approximates $\mathrm{Opt}_\Gamma(X,d)$ on any metric $d \in \mathrm{MET}(X)$ to within a factor of $O\left(\alpha(n)(\log n)^2\right)$.*

Proof. Let (X,d) be an n-point metric space and let $\mathscr{D}$ be the distribution over random ultrametrics ρ on X from Theorem 1 (which is computable in polynomial time, as follows directly from our proof of Theorem 1 in Section 2). In other words, $\rho(x,y) \ge d(x,y)$ for all $x,y \in X$ and

$$\max_{x\in X} \mathbb{E}_{\mathscr{D}} \max_{y\in X\setminus\{x\}} \frac{\rho(x,y)}{d(x,y)} \le C(\log n)^2.$$

Let $P \in 2^{X\times 2^X}$ be a clustering solution for which

$$\mathrm{Opt}_\Gamma(X,d) = \sum_{x\in X} \Gamma(x,d,P).$$

Using the monotonicity and homogeneity of Γ we see that

$$\mathrm{Opt}_\Gamma(X,\rho) \quad\le\quad \sum_{x\in X}\Gamma(x,\rho,P) \le \sum_{x\in X}\Gamma\left(x,\left[\max_{y\in X\setminus\{x\}}\frac{\rho(x,y)}{d(x,y)}\right]\cdot d, P\right) = \sum_{x\in X}\left[\max_{y\in X\setminus\{x\}}\frac{\rho(x,y)}{d(x,y)}\right]\cdot\Gamma(x,d,P).$$

Taking expectation we conclude that

$$\mathbb{E}_{\mathscr{D}}\mathrm{Opt}_\Gamma(X,\rho) \le \sum_{x\in X}\left(\mathbb{E}_{\mathscr{D}}\left[\max_{y\in X\setminus\{x\}}\frac{\rho(x,y)}{d(x,y)}\right]\right)\Gamma(x,d,P) \le C(\log n)^2\cdot\mathrm{Opt}_\Gamma(X,d).$$

Hence, with probability at least $\frac{1}{2}$ we have

$$\mathrm{Opt}_\Gamma(X,\rho) \le 2C(\log n)^2\cdot\mathrm{Opt}_\Gamma(X,d).$$

For such ρ compute a clustering solution $Q \in 2^{X\times 2^X}$ satisfying

$$\sum_{x\in X}\Gamma(x,\rho,Q) \le \alpha(n)\mathrm{Opt}_\Gamma(X,\rho) \le 2C\alpha(n)(\log n)^2\cdot\mathrm{Opt}_\Gamma(X,d).$$

Since $\rho \ge d$ it remains to use the monotonicity of Γ once more to deduce that

$$\sum_{x\in X}\Gamma(x,\rho,Q) \ge \sum_{x\in X}\Gamma(x,d,Q) \ge \mathrm{Opt}_\Gamma(X,d).$$

Thus Q is a $O\left(\alpha(n)(\log n)^2\right)$ approximate solution to the clustering problem on X with cost Γ. □

Due to Theorem 2 we see that the main difficulty in monotone clustering problems lies in the design of good approximation algorithms for them on ultrametrics. This is a generic reduction, and in many particular cases it might be possible use a case-specific analysis to improve the $O\left((\log n)^2\right)$ loss in the approximation factor. However, as a general reduction paradigm for clustering problems, Theorem 2 makes it clear why maximum gradient embeddings are so natural.

We will now present some specific applications of Theorem 2. Before doing so we would like to state at the outset that we do not know if it is the case that all monotone clustering problems can be well-approximated on ultrametrics. If this were true then Theorem 2 would yield an approximation algorithm for all monotone clustering problems. In Section 4.1 we will use a simple dynamic programming approach to design algorithms for certain monotone clustering problems. Our approach is quite flexible, and one can use it to give some additional conditions which show that a wide range of monotone clustering problems can be solved on ultrametrics. However carrying the analysis through in such generality seems to hide the simplicity of our ideas, so we prefer to analyze particular cases. In any case, we state the following natural question which remains open:

Open problem: Which monotone clustering problems have a polynomial time polylogarithmic approximation algorithm on ultrametrics?

We now describe some monotone clustering problems for which Theorem 2 yields the best known approximation algorithm.

• **Fault-tolerant k-median and facility location.** Fix $k \in \mathbb{N}$. The k-median problem is as follows. Given an n-point metric space (X, d_X), find $x_1, \ldots, x_k \in X$ that minimize the objective function

$$\sum_{x\in X} \min_{j\in\{x_1,\ldots,x_k\}} d_X(x, x_j). \tag{4}$$

This very natural and well studied problem can be easily cast as monotone clustering problem by defining $\Gamma(x, d, \{(x_1, C_1), \ldots, (x_m, C_m)\})$ to be ∞ if $m \neq k$, and otherwise

$$\Gamma(x, d, \{(x_1, C_1), \ldots, (x_m, C_m)\}) = \min_{j\in\{x_1,\ldots,x_k\}} d(x, x_j).$$

The linear structure of (4) makes it a prime example of a problem which can be approximated using Bartal's probabilistic embeddings. Indeed, the first non-trivial approximation algorithm for k-median clustering was obtained by Bartal in [4] (another such example is Min-Sum clustering— see [5]). Since then this problem has been investigated extensively: The first constant factor approximation for it was obtained in [14] using LP rounding, and the first combinatorial (primal-dual) constant-factor algorithm was obtained in [26]. In [2] an analysis of a natural local search heuristic yields the best known approximation factor for k-median clustering.

Here we study the following fault-tolerant version of the k-median problem. Let (X, d) be an n-point metric space and fix $k \in \mathbb{N}$. Assume that for every $x \in X$ we are given an integer $j(x) \in X$ (which we call the fault-tolerant parameter of x). Given $x_1, \ldots, x_k$ and $x \in X$ let $x_j^*(x; d)$ be the j-th closest point to x in $\{x_1, \ldots, x_k\}$. In other words, $\{x_j^*(x; d)\}_{j=1}^k$ is a re-ordering of $\{x_j\}_{j=1}^k$ such that $d(x, x_1^*(x; d)) \leq \cdots \leq d(x, x_k^*(x; d))$. Our goal is to minimize the objective function

$$\sum_{x\in X} d\left(x, x_{j(x)}^*(x; d)\right). \tag{5}$$

To understand (5) assume for the sake of simplicity that $j(x) = j$ for all $x \in X$. If $\{x_j\}_{j=1}^k$ minimize (5) and $j - 1$ of them are deleted (due to possible noise), then we are still ensured that on average every point in X is close to one of the x_j. In this sense the clustering problem in (5) is fault-tolerant. Observe that for $j = 1$ we return to the k-median clustering problem.

We remark that another fault-tolerant version of k-median clustering was introduced in [27]. In this problem we connect each point x in the metric space X to $j(x)$ centers, but the objective function is the sum over $x \in X$ of the sum of the distances from x to all the $j(x)$ centers. Once again, the linearity of the objective function seems to make the problem easier, and in [40] a constant factor approximation is achieved (this immediately implies that our version of fault-tolerant k-median clustering, i.e. the minimization of (5), has a $O\left(\max_{x\in X} j(x)\right)$ approximation algorithm). In particular, the LP that was previously used for k-median clustering naturally generalizes to this setting. This is not the case for our fault-tolerant version in (5). Moreover, the local search techniques for k-median clustering (see for example [2]) do not seem to be easily generalizable to the case $j > 1$, and in any case seem to require $n^{\Omega(j)}$ time, which is not polynomial even for moderate values of j.

Arguing as above in the case of k-median clustering we see that the fault-tolerant k-median clustering problem in (5) is a monotone clustering problem. In Section 4.1 we show that it can be solved exactly in polynomial time on ultrametrics. Thus, in combination with Theorem 2, we obtain a $O\left((\log n)^2\right)$ approximation algorithm for the minimization of (5) on general metrics.

Remark 1. Facility location type problems have been studied extensively since the 1960's— we refer to the book [35], and specifically to the chapter [16], for a discussion of such problems. The uncapacitated metric facility location problem is closely related to k-median problem (indeed k-median can be reduced to it via Lagrangian relaxation— see [26]), and has been studied extensively in recent years (see [13, 21, 25, 26, 28, 39]). In the context of (5) we can also consider the following fault-tolerant version of the facility location problem. Assume in addition that we are given non-negative facility costs $\{f_x\}_{x\in X}$. Then the goal is to minimize over all $x_1, \ldots, x_k \in X$ the objective function

$$\sum_{j=1}^{k} f_{x_j} + \sum_{x\in X} d\left(x, x^*_{j(x)}(x; d)\right). \tag{6}$$

The case $j(x) \equiv 1$ reduces to the classical un-capacitated metric facility location problem. The techniques presented here can be easily generalized to yield a $O\left((\log n)^2\right)$ approximation algorithm for the minimization of (6) as well.

• $\Sigma\ell_p$ **clustering.** Another problem which illustrates the usefulness of Theorem 2 is the $\Sigma\ell_p$ clustering problem which we now describe. Our argument for this problem is quite general, and it applies to more cost functions, but it is beneficial to concentrate on a concrete example. For $p \in [0, \infty]$ the $\Sigma\ell_p$ clustering problem is as follows: For a metric space (X, d) and $k \in \mathbb{N}$ the goal is to find $x_1, \ldots, x_k \in X$ and a partition of X into k sets $C_1, \ldots, C_k \subseteq X$ which minimize the objective function

$$\sum_{j=1}^{k} \left(\sum_{x\in C_j} d(x, x_j)^p \right)^{1/p}. \tag{7}$$

When $p = 1$ this becomes the k-median problem, and when $p = \infty$ this is the "sum of the cluster radii" problem, which has been studied in [15]. In both of these extreme cases there is a constant factor approximation algorithm known, so we automatically get a $O\left(\min\{n^{1/p}, n^{1-1/p}\}\right)$ approximation algorithm for (7). Here we shall use the framework of Theorem 2 to give a $O\left((\log n)^2\right)$ approximation algorithm for this problem for general p.

Observe that the $\Sigma\ell_p$ clustering problems are monotone clustering problems. Indeed, all we need to do is define $\Gamma(x, d, \{(x_1, C_1), \ldots, (x_m, C_m)\})$ to be ∞ if $\{C_1, \ldots, C_m\}$ is not a partition of X or $m \neq k$. Otherwise set $\Gamma(x, d, \{(x_1, C_1), \ldots, (x_k, C_k)\}) = 0$ if $x \notin \{x_1, \ldots, x_k\}$ and for $j \in \{1, \ldots, k\}$,

$$\Gamma(x_j, d, \{(x_1, C_1), \ldots, (x_k, C_k)\}) = \left(\sum_{x \in C_j} d(x, x_j)^p\right)^{1/p}.$$

This definition clearly makes Γ a homogeneous monotone clustering cost function. The following lemma, combined with Theorem 2, therefore implies that the $\Sigma\ell_p$ clustering problem has a $O\left((\log n)^2\right)$ approximation algorithm.

Lemma 3. *The $\Sigma\ell_p$ clustering problem has a constant factor polynomial time approximation algorithm (even a FPTAS) on ultrametrics.*

Lemma 3 will be proved via dynamic programming in Section 4.1.

2 Proof of the upper bound in Theorem 1

We start by recalling some terminology and results concerning random partitions of metric spaces. Given a partition $\mathscr{P}$ of a finite metric space (X, d_X) and $x \in X$ we denote by $\mathscr{P}(x)$ the unique element of $\mathscr{P}$ to which x belongs. For $\Delta > 0$ the partition $\mathscr{P}$ is said to be Δ-bounded if for every $x \in X$ we have $\mathrm{diam}(\mathscr{P}(x)) \leq \Delta$. We also fix a positive measure μ on X. The following fundamental result is due to [19] when μ is the uniform measure on X. The case of general measures was observed in [29, 32], and the specific numerical constants used below are taken from [34].

Lemma 4. *For every $\Delta > 0$ there exists a distribution over Δ-bounded partitions $\mathscr{P}$ of X such that for every $x \in X$ and every $0 < t \leq \Delta/8$,*

$$\Pr\left[B_X(x, t) \nsubseteq \mathscr{P}(x)\right] \leq \frac{16t}{\Delta} \cdot \log \frac{\mu(B_X(x, \Delta))}{\mu(B_X(x, \Delta/8))}. \tag{8}$$

We also recall the notion of a *quotient of a metric space* (see [9, 20, 33]). Let $\mathscr{W} = \{W_1, \ldots, W_m\}$ be a partition of X. For $W, W' \in \mathscr{W}$ write $d_X(W, W') = \min\{d_X(x, y) : \ x \in W, \ y \in W'\}$. The quotient metric space $(X/\mathscr{W}, d_{X/\mathscr{W}})$ is define as follows. As a set $X/\mathscr{W}$ coincides with $\mathscr{W}$. The metric $d_{X/\mathscr{W}}$ is the maximal metric on $\mathscr{W}$ which is majorized by $d_X(\cdot, \cdot)$. In other words, for $W, W' \in \mathscr{W}$,

$$d_{X/\mathscr{W}}(W, W') = \min\left\{\sum_{j=1}^{m-1} d_X(V_{j-1}, V_j) : \ V_0, \ldots, V_{m-1} \in \mathscr{W}, \ V_0 = W, \ V_{m-1} = W'\right\}.$$

The following lemma is a well known "quotient version" of Lemma 4. The argument below is known to experts, and appeared in various guises in several references— see for example [3, 24, 34]. Since we couldn't locate the formulation that we need in the literature, we include a proof here.

Lemma 5. *Let (X, d_X) be an n-point metric space and $\Delta > 0$. Then there exists a distribution over Δ-bounded partitions $\mathscr{P}$ of X such that for every $x, y \in X$, if $d_X(x, y) \leq \frac{\Delta}{2n^2}$ then $\mathscr{P}(x) = \mathscr{P}(y)$, and for every $x \in X$ and $0 < t \leq \Delta/16$,*

$$\Pr\left[B_X(x,t) \not\subseteq \mathscr{P}(x)\right] \leq \frac{32t}{\Delta} \cdot \log \frac{\mu(B_X(x,\Delta))}{\mu(B_X(x,\Delta/16))}.$$

Proof. Let ρ be an ultrametric on X such that $d_X(x, y) \leq \rho(x, y) \leq n d_X(x, y)$ for every $x, y \in X$. The simple construction of such a metric is contained in Lemma 3.6 in [6]. Define an equivalence relation on X by $x \sim y$ if $\rho(x, y) \leq \frac{\Delta}{2n}$. This is an equivalence relation since ρ is an ultrametric. Let $\mathscr{W} = \{W_1, \ldots, W_m\}$ be the equivalence classes of this relation, and consider the quotient metric space $X/\mathscr{W}$. We also denote by $\pi : X \to \mathscr{W}$ the induced quotient map, i.e. for $x \in W_j$, $\pi(x) = W_j$. Let $\mu \circ \pi^{-1}$ be the push-forward of the measure μ under π, i.e. $\mu \circ \pi^{-1}$ is the measure on $\mathscr{W}$ given for $W \in \mathscr{W}$ by $\mu \circ \pi^{-1}(W) = \mu(\pi^{-1}(W))$. Observe that for every $x, y \in X$,

$$d_X(x,y) - \frac{\Delta}{2} \leq d_{X/\mathscr{W}}(\pi(x), \pi(y)) \leq d_X(x,y). \tag{9}$$

Indeed, the upper bound in (9) is immediate from the definition of a quotient metric. The lower bound in (9) is proved as follows. There are points $x = x_0, x_1, \ldots, x_k = y$ in X such that $d_{X/\mathscr{W}}(\pi(x), \pi(y)) = \sum_{j=1}^k d_X(\pi(x_{j-1}), \pi(x_j))$. For $j \in \{1, \ldots, k\}$ let $a_j \in \pi(x_{j-1})$ and $b_j \in \pi(x_j)$ be such that $d_X(a_j, b_j) = d_X(\pi(x_{j-1}), \pi(x_j))$. Then, since $k \leq n-1$ and for all $z \in X$ we have $\operatorname{diam}(\pi(z)) = \max_{a,b\in\pi(z)} d_X(a, b) \leq \max_{a,b\in\pi(z)} \rho(a, b) \leq \frac{\Delta}{2n}$, we get that

$$\begin{aligned} d_X(x,y) &\leq d_X(x,a_1) + \sum_{j=1}^{k} d_X(a_j,b_j) + \sum_{j=1}^{k-1} d_X(b_j,a_{j+1}) + d_X(b_k,y) \\ &\leq \frac{\Delta}{2n} + d_{X/\mathscr{W}}(\pi(x),\pi(y)) + (n-2)\frac{\Delta}{2n} + \frac{\Delta}{2n}, \end{aligned}$$

implying the lower bound in (9).

Let $\mathscr{Q}$ be a distribution over $\Delta/2$-bounded partitions of $X/\mathscr{W}$ such that for every $W \in \mathscr{W}$ and every $0 < t \leq \Delta/16$ we have

$$\Pr\left[B_{X/\mathscr{W}}(W,t) \not\subseteq \mathscr{Q}(W)\right] \leq \frac{32t}{\Delta} \cdot \log \frac{\mu\circ\pi^{-1}(B_{X/\mathscr{W}}(W,\Delta/2))}{\mu\circ\pi^{-1}(B_{X/\mathscr{W}}(W,\Delta/16))}. \tag{10}$$

The existence of $\mathscr{Q}$ follows from Lemma 4. Let $\mathscr{P}$ be the pull-back of $\mathscr{Q}$ under π, i.e. $\mathscr{P}$ is the partition of X given by $\mathscr{P} = \{\pi^{-1}(A) : A \in \mathscr{Q}\}$. Note that (9) implies that for every $x \in X$ we have $\pi^{-1}\left(B_{X/\mathscr{W}}(\pi(x), \Delta/2)\right) \subseteq B_X(x, \Delta)$ and for every $t > 0$, $\pi^{-1}\left(B_{X/\mathscr{W}}(\pi(x), t)\right) \supseteq B_X(x, t)$. Thus (10) implies that for every $x \in X$ and $0 < t \leq \Delta/16$,

$$\Pr\left[B_X(x,t) \not\subseteq \mathscr{P}(x)\right] \leq \Pr\left[B_{X/\mathscr{W}}(\pi(x),t) \not\subseteq \mathscr{Q}(\pi(x))\right] \leq \frac{32t}{\Delta} \cdot \log \frac{\mu(B_X(x,\Delta))}{\mu(B_X(x,\Delta/16))}.$$

It remains to note that (9) implies that $\mathscr{P}$ is Δ-bounded and if $d_X(x, y) \leq \frac{\Delta}{2n^2}$ then $\rho(x, y) \leq \frac{\Delta}{2n}$. This means that $\pi(x) = \pi(y)$, so that $\mathscr{P}(x) = \mathscr{P}(y)$. □

Proof of the upper bound in Theorem 1. For every $k \in \mathbb{Z}$ let $\mathscr{P}_k$ be the distribution over partitions of X from Lemma 5 with $\Delta = 16^k$, where μ is the counting measure on X (we assume in what follows that the distributions $\{\mathscr{P}_k\}_{k\in\mathbb{Z}}$ are independent). For $x, y \in X$ let k be the largest integer for which $\mathscr{P}_k(x) \neq \mathscr{P}_k(y)$ (such a k must exists since for small enough k we have $\mathscr{P}_k(z) = \{z\}$ for all $z \in X$). Denote $\rho(x, y) = 16^{k+1}$. Then ρ is a (random) ultrametric on X. Indeed, if $x, y, z \in X$ and $\rho(x, y) = 16^{k+1}$ then $\mathscr{P}_k(x) \neq \mathscr{P}_k(y)$. It follows that either $\mathscr{P}_k(z) \neq \mathscr{P}_k(x)$ or $\mathscr{P}_k(z) \neq \mathscr{P}_k(y)$. Thus by the definition of ρ we have that $\max\{\rho(x, z), \rho(y, z)\} \geq \rho(x, y)$. Note also that if $\rho(x, y) = 16^{k+1}$ then $\mathscr{P}_{k+1}(x) = \mathscr{P}_{k+1}(y)$, so that $d_X(x, y) \leq \operatorname{diam}(\mathscr{P}(x)) \leq 16^{k+1} = \rho(x, y)$. It follows that the identity mapping on X is a random non-contractive embedding of X into the ultrametric (X, ρ). Finally, since whenever $d_X(x, y) \leq \frac{16^k}{2n^2}$ we have $\mathscr{P}_k(x) = \mathscr{P}_k(y)$, we are ensured that $\rho(x, y) \leq 32n^2 d_X(x, y)$ for every $x, y \in X$.

Denote for $x \in X$ and $i \in \mathbb{Z}$, $A_i(x) = B_X(x, 16^i) \setminus B_X(x, 16^{i-1})$. For every $j \in \mathbb{N}$ and $k \in \mathbb{Z}$ we claim that the following inclusion of events holds true

$$\left\{\exists y \in A_{k-j}(x)\ \ 16^{j+2} > \frac{\rho(x, y)}{d_X(x, y)} \geq 16^{j+1}\right\} \subseteq \left\{B_X(x, 16^{k-j}) \not\subseteq \mathscr{P}_k(x)\right\}. \tag{11}$$

Indeed, assume for the sake of contradiction that (11) fails. Then it is possible that $B_X(x, 16^{k-j}) \subseteq \mathscr{P}_k(x)$, yet there exists $y \in A_{k-j}(x)$ satisfying $16^{j+2} d_X(x, y) > \rho(x, y) \geq 16^{j+1} d_X(x, y)$. Let $s \in \mathbb{N}$ be such that $\rho(x, y) = 16^{s+1}$. Since $y \in B_X(x, 16^{k-j})$ we know that $16^{s+1} = \rho(x, y) < 16^{j+2} \cdot 16^{k-j} = 16^{k+2}$. Thus $s \leq k$. Since $y \in B_X(x, 16^{k-j}) \subseteq \mathscr{P}_k(x)$ we also know that $\mathscr{P}_k(x) = \mathscr{P}_k(y)$, which by the definition of ρ implies that $s \neq k$. Thus $s \leq k - 1$. But since $y \notin B_X(x, 16^{k-j-1})$ we have that $\rho(x, y) = 16^{s+1} \leq 16^k < 16^{j+1} d_X(x, y)$, which is a contradiction to our assumption on y.

Using (11) we have

$$\Pr\left[\exists y \in A_{k-j}(x)\ \ 16^{j+2} > \frac{\rho(x, y)}{d_X(x, y)} \geq 16^{j+1}\right] \leq \Pr\left[B_X(x, 16^{k-j}) \not\subseteq \mathscr{P}_k(x)\right] \leq \frac{32}{16^j} \cdot \log \frac{|B_X(x, 16^k)|}{|B_X(x, 16^{k-1})|}.$$

Thus, since $X = \bigcup_{i\in\mathbb{Z}} A_i(x)$, we see that

$$\begin{aligned}\Pr\left[16^{j+1} > \max_{y\in X\setminus\{x\}} \frac{\rho(x, y)}{d_X(x, y)} \geq 16^j\right] &\leq \Pr\left[\bigcup_{i\in\mathbb{Z}} \left\{\exists y \in A_i(x)\ \ 16^{j+1} > \frac{\rho(x, y)}{d_X(x, y)} \geq 16^j\right\}\right] \\ &\leq \sum_{i\in\mathbb{Z}} \Pr\left[\exists y \in A_i(x)\ \ 16^{j+1} > \frac{\rho(x, y)}{d_X(x, y)} \geq 16^j\right] \leq \sum_{i\in\mathbb{Z}} \frac{32}{16^{j-1}} \cdot \log \frac{|B_X(x, 16^{i+j-1})|}{|B_X(x, 16^{i+j-2})|} \leq \frac{512}{16^j} \cdot \log n.\end{aligned}$$

Hence, using the a priori bound $\rho(x, y) \leq 32n^2 d_X(x, y)$, it follows that

$$\begin{aligned}\mathbb{E}\left[\max_{y\in X\setminus\{x\}} \frac{\rho(x, y)}{d_X(x, y)}\right] &\leq 16 \Pr\left[\max_{y\in X\setminus\{x\}} \frac{\rho(x, y)}{d_X(x, y)} \leq 16\right] + \sum_{j=1}^{\lfloor \log_{16}(32n^2)\rfloor} 16^{j+1} \Pr\left[16^{j+1} > \max_{y\in X\setminus\{x\}} \frac{\rho(x, y)}{d_X(x, y)} \geq 16^j\right] \\ &\leq 16 + \sum_{j=1}^{\lfloor \log_{16}(32n^2)\rfloor} 16^{j+1} \cdot \frac{512}{16^j} \cdot \log n = O\left(1 + (\log n)^2\right).\end{aligned}$$

This completes the proof of the upper bound in Theorem 1. □

Remark 2. The above argument also shows that for every n-point metric space (X, d_X) there exists a distribution over non-contractive embeddings into ultrametrics $f : X \to U$ such that

$$\mathbb{E}_{\mathscr{D}} |\nabla f(x)|_\infty = O\left(1 + (\log n) \log \Phi(X)\right),$$

where $\Phi(X)$ is the aspect ratio of X, which is defined by

$$\Phi(X) = \frac{\operatorname{diam} X}{\min_{\substack{x,y\in X\\ x\neq y}} d_X(x,y)} = \frac{\max_{x,y\in X} d_X(x,y)}{\min_{\substack{x,y\in X\\ x\neq y}} d_X(x,y)}.$$

3 Tight lower bounds for cycles, paths, and diamond graphs

As mentioned in the introduction, the metrics Y_n in Theorem 1 are the diamond graphs of Newman and Rabinovich [37], which will be defined presently. Before passing to this more complicated (and strongest) lower bound, we will analyze the simpler examples of cycles and paths, which are of independent interest.

Let C_n, $n > 3$, be the unweighted path on n-vertices. We will identify C_n with the group $\mathbb{Z}_n$ of integers modulo n. We first observe that in this special case the upper bound in Theorem 1 can be improved to $O(\log n)$. This is achieved by using Karp's embedding of the cycle into spanning paths— we simply choose an edge of C_n uniformly at random and delete it. Let $f : C_n \to \mathbb{Z}$ be the randomized embedding thus obtained, which is clearly non-contractive.

Karp noted that it is easy to see that as a probabilistic embedding into trees f has distortion at most 2. We will now show that as a maximum gradient embedding, f has distortion $\Theta(\log n)$. Indeed, fix $x \in C_n$, and denote the deleted edge by $\{a, a+1\}$. Assume that $d_{C_n}(x,a) = t \le n/2 - 1$. Then the distance from $a+1$ to x changed from $t+1$ in C_n to $n-t-1$ in the path. It is also easy to see that this is where the maximum gradient is attained. Thus

$$\mathbb{E}|\nabla f(x)|_\infty \approx \frac{2}{n}\sum_{0\le t\le n/2}\frac{n-t-1}{t+1} = \Theta(\log n).$$

We will now show that any maximum gradient embedding of C_n into a distribution over trees incurs distortion $\Omega(\log n)$. For this purpose we will use the following lemma from [38].

Lemma 6. *For any tree metric T, and any non-contractive embedding $g : C_n \to T$, there exists an edge $(x, x+1)$ of C_n such that $d_T(g(x), g(x+1)) \ge \frac{n}{3} - 1$.*

Now, let $\mathscr{D}$ be a distribution over non-contractive embeddings of C_n into trees $f : C_n \to T$. By Lemma 6 we know that there exists $x \in C_n$ such that $d_T(f(x), f(x+1)) \ge \frac{n-3}{3}$. Thus for every $y \in C_n$ we have that $\max\{d_T(f(y), f(x)), d_T(f(y), f(x+1))\} \ge \frac{n-3}{6}$. On the other hand $\max\{d_{C_n}(y,x), d_{C_n}(y, x+1)\} \le d_{C_n}(x,y)+1$. It follows that

$$|\nabla f(y)|_\infty \ge \frac{n-3}{6d_{C_n}(x,y)+6}.$$

Summing this inequality over $y \in C_n$ we see that

$$\sum_{y\in C_n}|\nabla f(y)|_\infty \ge \sum_{0\le k\le n/2}\frac{n-3}{6k+6} = \Omega(n\log n).$$

Thus

$$\max_{y\in C_n}\mathbb{E}_{\mathscr{D}}|\nabla f(y)|_\infty \ge \frac{1}{n}\sum_{y\in C_n}\mathbb{E}_{\mathscr{D}}|\nabla f(y)|_\infty = \Omega(\log n),$$

as required.

We will now deal with the more complicated case of maximum gradient embeddings of the unweighted path on n-vertices, which we denote by P_n, into ultrametrics. The following proposition shows that Theorem 1 is optimal when one considers embeddings into ultrametrics. This is weaker than the lower bound in Theorem 1, which deals with embeddings into arbitrary trees (note that P_n is a tree).

Proposition 7. *Let $\mathscr{D}$ be a distribution over non-contractive embeddings of P_n into ultrametrics $f : P_n \to U$. Then there exists $x \in P_n$ such that $\mathbb{E}_{\mathscr{D}}|\nabla f(x)|_\infty = \Omega((\log n)^2)$.*

Before proving Proposition 7 we record the following numerical inequalities.

Lemma 8. *The following elementary inequalities hold true:*

1. *For every $a, b \in \{0, 1, 2, \dots\}$,*
$$a(\log a)^2 + b(\log b)^2 \geq (a+b)\left(\log(a+b)\right)^2 - 2\left[1 + \log\left(\frac{a+b}{a}\right)\right] a \log(a+b).$$
2. *For every $x \geq 1$, $(1 + \log x)\log x \leq 4\sqrt{x}$.*

Proof. The first inequality is trivial if $a = 0$ or $b = 0$, so assume that $a, b \geq 1$. Denote for $t \geq 0$, $\psi(t) = t(\log t)^2$. Then

$$\begin{aligned}
(a+b)\left(\log(a+b)\right)^2 - b(\log b)^2 &= \int_b^{a+b} \psi'(t)dt \\
&= \int_b^{a+b} \left[(\log t)^2 + 2\log t\right] dt \\
&\leq a\left(\log(a+b)\right)^2 + 2a\log(a+b) \\
&= a(\log a)^2 + a\left[\log(a+b) + \log a\right] \cdot \log\left(\frac{a+b}{a}\right) + 2a\log(a+b) \\
&\leq a(\log a)^2 + 2\left[1 + \log\left(\frac{a+b}{a}\right)\right] a\log(a+b),
\end{aligned}$$

proving the first assertion in Lemma 8.

The second assertion in Claim 8 follows from the inequality $\log x \leq 2\sqrt[4]{x} - 1$, which is true since the minimum of the function $y \mapsto 2\sqrt[4]{y} - 1 - \log y$, which is attained at $y = 16$, is positive. $\square$

Proof of Proposition 7. We think of P_n as the interval of integers $I = \{0, \dots, n-1\} \subseteq \mathbb{R}$. Arguing the same as in the case of the cycle C_n, it is enough to prove that if (U, d_U) is an ultrametric and $f : P_n \to U$ is non-contractive then

$$\frac{1}{n}\sum_{x=0}^{n-1} |\nabla f(x)|_\infty \geq c(\log n)^2, \tag{12}$$

where $c > 0$ is a universal constant.

Given a sub-interval $J = \{a, a+1, \dots, a+t\} \subseteq \{0, \dots, n-1\}$ let m_J be the largest point $m \in \{a+1, \dots, a+t\}$ for which $d_U(f(m-1), f(m)) = \|f|_J\|_{\mathrm{Lip}} = \max_{1 \leq i \leq t} d_U(f(a+i-1), f(a+i))$ (if $t = 0$ then we set $m_J = a$). Since the distortion of J in any embedding into an ultrametric is at least $|J| - 1$ (see Lemma 2.4 in [33]), we know that $d_U(f(m_J - 1), f(m_J)) \geq t = |J| - 1$. We shall denote in what follows J_s to be the shorter of the two intervals $\{a, a+1, \dots, m_J - 1\}$ and $\{m_J, \dots, a+t\}$ (breaking ties arbitrarily), and J_b will denote the longer of these two intervals (when $|J| = 1$ we use the convention $J_s = J_b$). Thus $J = J_s \cup J_b$ and $|J_s| \leq |J_b|$. Finally, let x_J be the point in J_s which is closest to J_b (so that $x_J \in \{m_J, m_{J-1}\}$).

We define a function $g_J : J \to \mathbb{R}$ inductively as follows. If $1 \le |J_s| \le \sqrt{|J|}$ then

$$g_J(x) = \begin{cases} g_{J_s}(x) & \text{if } x \in J_s \setminus \{x_J\}, \\ \frac{1}{8}\left[1 + \log\left(\frac{|J|}{|J_s|}\right)\right] |J_s| \log |J| & \text{if } x = x_J, \\ g_{J_b}(x) & \text{if } x \in J_b. \end{cases} \tag{13}$$

If, on the other hand, $|J_s| > \sqrt{|J|}$ then

$$g_J(x) = \begin{cases} g_{J_s}(x) & \text{if } x \in J_s \text{ and } |x - x_J| > \sqrt[4]{|J_s|}, \\ \frac{|J|-1}{|x-x_J|+1} & \text{if } x \in J_s \text{ and } |x - x_J| \le \sqrt[4]{|J_s|}, \\ g_{J_b}(x) & \text{if } x \in J_b. \end{cases} \tag{14}$$

The following claim summarizes the crucial properties of the these mappings. Recall that we are using the notation $I = \{0, \dots, n-1\}$.

Claim 9. *The following assertions hold true for every sub-interval $J \subseteq I$.*

1. *For every $x \in J$ we have $g_J(x) \le |\nabla(f|_J)(x)|_\infty = \max_{y \in J \setminus \{x\}} \frac{d_U(f(x), f(y))}{|x-y|}$.*
2. *For every $x \in J$, $g_J(x) \le |J| - 1$.*
3. *If $|J_s| \ge \sqrt{J}$ and $|x - x_J| \le \sqrt[4]{|J_s|}$ then $g_{J_s}(x) \le 4\sqrt{|J_s|}$.*

Proof. The proofs of all of the assertion in Claim 9 will be by induction on J. To prove the first assertion assume first that $1 \le |J_s| \le \sqrt{|J|}$. From the recursive definition in (13) it follows that we should show that $\frac{1}{8}\left[1 + \log\left(\frac{|J|}{|J_s|}\right)\right] |J_s| \log |J| \le |\nabla(f|_J)(x_J)|_\infty$. Since $x_J \in \{m_J - 1, m_J\}$ the definition of m_J implies that $|\nabla(f|_J)(x_J)|\infty \ge |J| - 1$. Thus it is enough to show that $\frac{1}{8}(1 + \log |J|)\sqrt{|J|} \log |J| \le |J| - 1$, which follows from the second assertion in Lemma 8. If, on the other hand, $|J_s| > \sqrt{|J|}$ then from the recursive definition in (14) it follows that it is enough to show that for every $x \in J_s$ we have $\frac{|J|-1}{|x-x_J|+1} \le |\nabla(f|_J)(x)|_\infty$. But since U is an ultrametric we know that

$$|J| - 1 \le d_U(f(m_J - 1), f(m_J)) \le \max\{d_U(f(x), f(m_J - 1)), d_U(f(x), f(m_J))\},$$

which implies the required lower bound on $|\nabla(f|_J)(x)|_\infty$ since $x_J \in \{m_J - 1, m_J\}$. The second assertion in Claim 9 is proved similarly.

It remains to prove the third assertion in Lemma 9. Let $K \subseteq J_s$ be the sub-interval of J_s in which the value of $g_{J_s}(x)$ was first set. In other words, $K \subseteq J_s$ is the smallest interval for which $x \in K_s$ and $g_K(x) = g_{J_s}(x)$. It follows in particular that $|x - x_K| \le \sqrt[4]{|K_s|}$. Also, by construction it is always the case that either K_s or K_b is contained in the interval $[\min\{x_K, x_J\}, \max\{x_K, x_J\}]$. Since K_s is shorter than K_b we are assured that

$$|K_s| \le |x_K - x_J| \le |x_K - x| + |x - x_J| \le \sqrt[4]{|K_s|} + \sqrt[4]{|J_s|} \le 2\sqrt[4]{|J_s|}. \tag{15}$$

If $|K_s| \le \sqrt{|K|}$ then necessarily $x = x_K$ and $g_K(x)$ was determined by the second line in (13). Hence

$$g_{J_s}(x) = g_K(x) = \frac{1}{8}\left[1 + \log\left(\frac{|K|}{|K_s|}\right)\right] |K_s| \log |K| \le \frac{1}{4}\left[1 + \log |J_s|\right] \sqrt[4]{|J_s|} \log |J_s| \le 4\sqrt{|J_s|}, \tag{16}$$

where we used (15) and the last inequality in (16) follows from the second assertion of Lemma 8.

Otherwise $|K_s| > \sqrt{|K|}$ and $g_K(x)$ was determined by the second line in (14), i.e.

$$g_{J_s}(x) = g_K(x) = \frac{|K|-1}{|x-x_K|+1} < |K| < |K_s|^2 \leq 4\sqrt{|J_s|},$$

where we used (15). This completes the proof of Claim 9. □

With Claim 9 at hand we are in position to conclude the proof of Proposition 7. We will prove by induction on $|J|$ that

$$\sum_{x\in J} g_J(x) \geq c|J|(\log|J|)^2. \tag{17}$$

This will prove (12), and hence imply Proposition 7, since by the first assertion of Claim 9 we get that

$$\sum_{x=0}^{n-1} |\nabla f(x)|_\infty \geq \sum_{x\in I} g_I(x) \geq cn(\log n)^2.$$

Inequality (17) trivially holds true with small enough constant c if $|J| \leq 2^{60}$, so assume that $|J| > 2^{60}$. To prove (17) we distinguish between two cases. If $|J_s| \leq \sqrt{|J|}$ then since $g_{J_s}(x_J) \leq |J_s|$ (by the second assertion in Claim 9) we see by induction that

$$\begin{aligned}
\sum_{x\in J} g_J(x) &= \sum_{x\in J_s} g_{J_s}(x) + \sum_{x\in J_b} g_{J_b}(x) + g_J(x_J) - g_{J_s}(x_J) \\
&> c\left(|J_s|(\log|J_s|)^2 + |J_b|(\log|J_b|)^2\right) + 2\left[1+\log\left(\frac{|J|}{|J_s|}\right)\right]|J_s|\log|J| - |J_s| && (18)\\
&\geq c|J|(\log|J|)^2 - 2c\left[1+\log\left(\frac{|J|}{|J_s|}\right)\right]|J_s|\log|J| + \left[1+\log\left(\frac{|J|}{|J_s|}\right)\right]|J_s|\log|J| && (19)\\
&\geq c|J|(\log|J|)^2, && (20)
\end{aligned}$$

where in (18) we used the inductive hypothesis and the inductive definition in (13) , in (19) we used Lemma 8, and (20) holds for $c \leq \frac{1}{2}$.

On the other hand if $|J_s| > \sqrt{|J|}$ then

$$\begin{aligned}
\sum_{x\in J} g_J(x) &= \sum_{x\in J_s} g_{J_s}(x) + \sum_{x\in J_b} g_{J_b}(x) + \sum_{\substack{x\in J_s \\ |x-x_J|\leq \sqrt[4]{|J_s|}}} \left(\frac{|J|-1}{|x-x_J|+1} - g_{J_s}(x)\right) && (21)\\
&\geq c|J|(\log|J|)^2 - 2c\left[1+\log\left(\frac{|J|}{|J_s|}\right)\right]|J_s|\log|J| + \sum_{k=0}^{\lfloor \sqrt[4]{|J_s|}\rfloor} \frac{|J|-1}{k+1} - 8|J_s|^{3/4} && (22)\\
&\geq c|J|(\log|J|)^2 - 2c\left[1+\log\left(\frac{|J|}{|J_s|}\right)\right]|J_s|\log|J| + \frac{1}{4}(|J|-1)\log|J_s| - 8|J|^{3/4} \\
&\geq c|J|(\log|J|)^2 - 2c\left[1+\log\left(\frac{|J|}{|J_s|}\right)\right]|J_s|\log|J| + \frac{1}{8}(|J|-1)\log|J_s| && (23)\\
&\geq c|J|(\log|J|)^2, && (24)
\end{aligned}$$

where in (21) we used the inductive definition in (14), in (22) we used the inductive hypothesis, Lemma 8 and Claim 9, and inequalities (23) and (24) hold for $|J| > 2^{60}$ and small enough c, respectively, since $\frac{|J|}{2} \leq |J_s| > \sqrt{|J|}$. This completes the proof of Proposition 7. □

We now pass to the proof of the lower bound in Theorem 1 in its full strength, i.e. in the case of maximum gradient embeddings into trees. We start by describing the diamond graphs $\{G_k\}_{k=1}^{\infty}$, and a special labelling of them that we will use throughout the ensuing arguments. The first diamond graph G_1 is a cycle of length 4, and G_{k+1} is obtained from G_k by replacing each edge by a quadrilateral. Thus G_k has 4^k edges and $\frac{2\cdot 4^k+4}{3}$ vertices. As we have done before, the required lower bound on maximum gradient embeddings of G_k into trees will be proved if we show that for every tree T and every non-contractive embedding $f : G_k \to T$ we have

$$\frac{1}{4^k}\sum_{e\in E(G_k)}\sum_{x\in e}|\nabla f(x)|_\infty = \Omega\left(k^2\right). \tag{25}$$

Note that the inequality (25) is different from the inequalities that we proved in the case of the cycle and the path in that the weighting on the vertices of G_k that it induces is not uniform— high degree vertices get more weight in the average in the left-hand side of (25).

We will prove (25) by induction on k. In order to facilitate such an induction, we will first strengthen the inductive hypothesis. To this end we need to introduce a useful labelling of G_k. For $1 \le i \le k$ the graph G_k contains 4^{k-i} canonical copies of G_i, which we index by elements of $\{1,2,3,4\}^{k-i}$, and denote $\left\{G^{(k)}_{[\alpha]}\right\}_{\alpha\in\{1,2,3,4\}^{k-i}}$. These graphs are defined as follows—see Figures 1 and 2 for a schematic description.

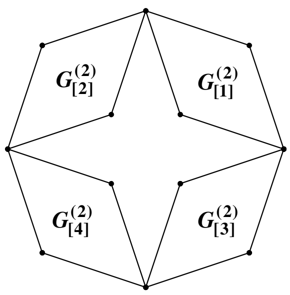

Figure 1: The graph G_2 and the labelling of the canonical copies of G_1 contained in it.

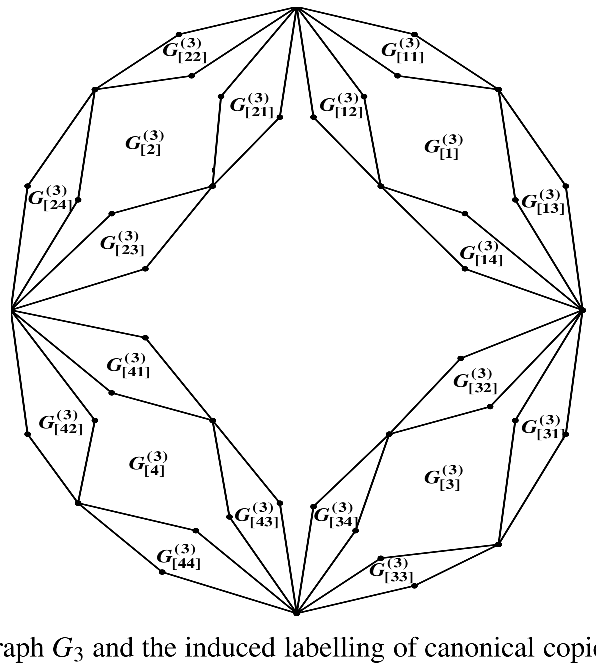

Figure 2: The graph G_3 and the induced labelling of canonical copies of G_1 and G_2.

Formally, we set $G^{(k)}_{[\emptyset]} = G_k$, and assume inductively that the canonical subgraphs of G_{k-1} have been defined. Let H_1, H_2, H_3, H_4 be the top-right, top-left, bottom-right and bottom-left copies of G_{k-1} in G_k, respectively. For $\alpha \in \{1,2,3,4\}^{k-1-i}$ and $j \in \{1,2,3,4\}$ we denote the copy of G_i in H_j corresponding to $G^{(k-1)}_{[\alpha]}$ by $G^{(k)}_{[j\alpha]}$.

For every $1 \le i \le k$ and $\alpha \in \{1,2,3,4\}^{k-i}$ let $T^{(k)}_{[\alpha]}, B^{(k)}_{[\alpha]}, L^{(k)}_{[\alpha]}, R^{(k)}_{[\alpha]}$ be the topmost, bottom-most, left-most, and right-most vertices of $G^{(k)}_{[\alpha]}$, respectively. We will construct inductively a set of simple cycles $\mathscr{C}_{[\alpha]}$ in $G^{(k)}_{[\alpha]}$ and for each $C \in \mathscr{C}_{[\alpha]}$ an edge $\varepsilon_C \in E\left(\mathscr{C}_{[\alpha]}\right)$, with the following properties.

1. The cycles in $\mathscr{C}_{[\alpha]}$ are edge-disjoint, and they all pass through the vertices $T^{(k)}_{[\alpha]}, B^{(k)}_{[\alpha]}, L^{(k)}_{[\alpha]}, R^{(k)}_{[\alpha]}$. There are 2^{i-1} cycles in $\mathscr{C}_{[\alpha]}$, and each of them contains 2^{i+1} edges. Thus in particular the cycles in $\mathscr{C}_{[\alpha]}$ form a disjoint cover of the edges in $G^{(k)}_{[\alpha]}$.

2. If $C \in \mathscr{C}_{[\alpha]}$ and $\varepsilon_C = \{x, y\}$ then $d_T(f(x), f(y)) \ge \frac{2^{i+1}}{3} - 1$.

3. Denote $E_{[\alpha]} = \{\varepsilon_C : C \in \mathscr{C}_{[\alpha]}\}$ and $\Delta_i = \bigcup_{\alpha\in\{1,2,3,4\}^{k-i}} E_{[\alpha]}$. The edges in Δ_i will be called the *designated edges* of level i. For $\alpha \in \{1,2,3,4\}^{k-i}$, $C \in \mathscr{C}_{[\alpha]}$ and $j < i$ let $\Delta_j(C) = \Delta_j \cap E(C)$ be the designated edges of level j on C. Then we require that each of the two paths $T^{(k)}_{[\alpha]} - L^{(k)}_{[\alpha]} - B^{(k)}_{[\alpha]}$ and $T^{(k)}_{[\alpha]} - R^{(k)}_{[\alpha]} - B^{(k)}_{[\alpha]}$ in C contain exactly 2^{i-j-1} edges from $\Delta_j(C)$.

The construction is done by induction on i. For $i = 1$ and $\alpha \in \{1,2,3,4\}^{k-1}$ we let $\mathscr{C}_{[\alpha]}$ contain only the 4-cycle $G^{(k)}_{[\alpha]}$ itself. Moreover by Lemma 6 there is and edge $\varepsilon_{G^{(k)}_{[\alpha]}} \in E\left(G^{(k)}_{[\alpha]}\right)$ such that if $\varepsilon_{G^{(k)}_{[\alpha]}} = \{x, y\}$ then $d_T(f(x), f(y)) \ge \frac{1}{3}$. This completes the construction for $i = 1$. Assuming we have completed the construction for $i - 1$ we construct the cycles at level i as follows. Fix arbitrary cycles $C_1 \in \mathscr{C}_{[1\alpha]}, C_2 \in \mathscr{C}_{[2\alpha]}, C_3 \in \mathscr{C}_{[3\alpha]}, C_4 \in \mathscr{C}_{[4\alpha]}$. We will use these four cycles to construct two cycles in $\mathscr{C}_{[\alpha]}$. The first one consists of the $T^{(k)}_{[\alpha]} - R^{(k)}_{[\alpha]}$ path in C_1 which contains the edge ε_{C_1}, the $R^{(k)}_{[\alpha]} - B^{(k)}_{[\alpha]}$ path in C_3 which does not contain the edge ε_{C_3}, the $B^{(k)}_{[\alpha]} - L^{(k)}_{[\alpha]}$ path in C_4 which contains the edge ε_{C_4}, and the $L^{(k)}_{[\alpha]} - T^{(k)}_{[\alpha]}$ path in C_2 which does not contain the edge ε_{C_2}. The remaining edges in $E(C_1) \cup E(C_2) \cup E(C_3) \cup E(C_4)$ constitute the second cycle that we extract from C_1, C_2, C_3, C_4. Continuing in this manner by choosing cycles from $\mathscr{C}_{[1\alpha]} \setminus \{C_1\}$, $\mathscr{C}_{[2\alpha]} \setminus \{C_2\}$, $\mathscr{C}_{[3\alpha]} \setminus \{C_3\}$, $\mathscr{C}_{[4\alpha]} \setminus \{C_4\}$ and repeating this procedure, and then continuing until we exhaust the cycles in $\mathscr{C}_{[1\alpha]} \cup \mathscr{C}_{[2\alpha]} \cup \mathscr{C}_{[3\alpha]} \cup \mathscr{C}_{[4\alpha]}$, we obtain the set of cycles $\mathscr{C}_{\alpha}$. For every $C \in \mathscr{C}_{\alpha}$ we then apply Lemma 6 to obtain an edge ε_C with the required property.

For each edge $e \in E(G_k)$ let $\alpha \in \{1,2,3,4\}^{k-i}$ be the unique multi-index such that $e \in E\left(G^{(k)}_{[\alpha]}\right)$. We denote by $C_i(e)$ the unique cycle in $\mathscr{C}_{[\alpha]}$ containing e. We will also denote $\widehat{e_i}(e) = \varepsilon_{C_i(e)}$. Finally we let $a_i(e) \in e$ and $b_i(e) \in \widehat{e_i}(e)$ be vertices such that

$$d_T(f(a_i(e)), f(b_i(e))) = \max_{\substack{a\in e\\ b\in\widehat{e_i}(e)}} d_T(f(a), f(b)).$$

Note that by the definition of $\widehat{e_i}(e)$ and the triangle inequality we are assured that

$$d_T(f(a_i(e)), f(b_i(e))) \ge \frac{1}{2}\left(\frac{2^{i+1}}{3} - 1\right) \ge \frac{2^i}{12}. \tag{26}$$

Recall that we plan to prove (25) by induction on k. Having done all of the above preparation, we are now in position to strengthen (25) so as to make the inductive argument easier. Given two edges $e, h \in G_k$ we write $e \frown_i h$ if both e, h are on the same canonical copy of G_i in G_k, $C_i(e) = C_i(h) = C$, and furthermore

e and h on the same side of C. In other words, $e \frown_i h$ if there is $\alpha \in \{1,2,3,4\}^{k-i}$ and $C \in \mathscr{C}_{[\alpha]}$ such that if we partition the edges of C into two disjoint $T^{(k)}_{[\alpha]} - B^{(k)}_{[\alpha]}$ paths, then e and h are on the same path.

Let $m \in \mathbb{N}$ be a universal constant that will be specified later. For every integer $\ell \leq k/m$ and any $\alpha \in \{1,2,3,4\}^{k-m\ell}$ define

$$L_\ell(\alpha) = \frac{1}{4^{m\ell}} \sum_{e \in E\left(G^{(k)}_{[\alpha]}\right)} \max_{\substack{i \in \{1,\dots,\ell\} \\ e \frown_{im} \widehat{e}_{im}(e)}} \frac{d_T(f(a_{im}(e)), f(b_{im}(e))) \wedge 2^{im}}{d_{G_k}(e, \widehat{e}_{im}(e)) + 1}.$$

We also write $L_\ell = \min_{\alpha \in \{1,2,3,4\}^{k-m\ell}} L_\ell(\alpha)$. We will prove that $L_\ell \geq L_{\ell-1} + c\ell$, where $c > 0$ is a universal constant. This will imply that for $\ell = \lfloor k/m \rfloor$ we have $L_\ell = \Omega(k^2)$ (since m is a universal constant). By simple arithmetic (25) follows.

Observe that for every $\alpha \in \{1,2,3,4\}^{k-m\ell}$ we have

$$\begin{aligned}
L_\ell(\alpha) &= \frac{1}{4^m} \sum_{\beta \in \{1,2,3,4\}^m} \frac{1}{4^{m(\ell-1)}} \sum_{e \in E\left(G^{(k)}_{[\beta\alpha]}\right)} \max_{\substack{i \in \{1,\dots,\ell\} \\ e \frown_{im} \widehat{e}_{im}(e)}} \frac{d_T(f(a_{im}(e)), f(b_{im}(e))) \wedge 2^{im}}{d_{G_k}(e, \widehat{e}_{im}(e)) + 1} \\
&= \frac{1}{4^m} \sum_{\beta \in \{1,2,3,4\}^m} \frac{1}{4^{m(\ell-1)}} \sum_{e \in E\left(G^{(k)}_{[\beta\alpha]}\right)} \max_{\substack{i \in \{1,\dots,\ell-1\} \\ e \frown_{im} \widehat{e}_{im}(e)}} \frac{d_T(f(a_{im}(e)), f(b_{im}(e))) \wedge 2^{im}}{d_{G_k}(e, \widehat{e}_{im}(e)) + 1} \\
&\quad + \frac{1}{4^{m\ell}} \sum_{e \in E\left(G^{(k)}_{[\alpha]}\right)} \max\left\{0, \frac{d_T(f(a_{\ell m}(e)), f(b_{\ell m}(e))) \wedge 2^{\ell m}}{d_{G_k}(e, \widehat{e}_{\ell m}(e)) + 1} \cdot \mathbf{1}_{\{e \frown_{\ell m} \widehat{e}_{\ell m}(e)\}} \right. \\
&\qquad\qquad \left. - \max_{\substack{i \in \{1,\dots,\ell-1\} \\ e \frown_{im} \widehat{e}_{im}(e)}} \frac{d_T(f(a_{im}(e)), f(b_{im}(e))) \wedge 2^{im}}{d_{G_k}(e, \widehat{e}_{im}(e)) + 1} \right\} \\
&= \frac{1}{4^m} \sum_{\beta \in \{1,2,3,4\}^m} L_{\ell-1}(\beta\alpha) \\
&\quad + \frac{1}{4^{m\ell}} \sum_{e \in E\left(G^{(k)}_{[\alpha]}\right)} \max\left\{0, \frac{d_T(f(a_{\ell m}(e)), f(b_{\ell m}(e))) \wedge 2^{\ell m}}{d_{G_k}(e, \widehat{e}_{\ell m}(e)) + 1} \cdot \mathbf{1}_{\{e \frown_{\ell m} \widehat{e}_{\ell m}(e)\}} \right. \\
&\qquad\qquad \left. - \max_{\substack{i \in \{1,\dots,\ell-1\} \\ e \frown_{im} \widehat{e}_{im}(e)}} \frac{d_T(f(a_{im}(e)), f(b_{im}(e))) \wedge 2^{im}}{d_{G_k}(e, \widehat{e}_{im}(e)) + 1} \right\} \\
&\geq L_{\ell-1} + \frac{1}{4^{m\ell}} \sum_{e \in E\left(G^{(k)}_{[\alpha]}\right)} \max\left\{0, \frac{d_T(f(a_{\ell m}(e)), f(b_{\ell m}(e))) \wedge 2^{\ell m}}{d_{G_k}(e, \widehat{e}_{\ell m}(e)) + 1} \cdot \mathbf{1}_{\{e \frown_{\ell m} \widehat{e}_{\ell m}(e)\}} \right. \\
&\qquad\qquad \left. - \max_{\substack{i \in \{1,\dots,\ell-1\} \\ e \frown_{im} \widehat{e}_{im}(e)}} \frac{d_T(f(a_{im}(e)), f(b_{im}(e))) \wedge 2^{im}}{d_{G_k}(e, \widehat{e}_{im}(e)) + 1} \right\}.
\end{aligned}$$

Thus it is enough to show that

$$A \stackrel{\text{def}}{=} \frac{1}{4^{m\ell}} \sum_{e \in E\left(G^{(k)}_{[\alpha]}\right)} \max\left\{0, \frac{d_T(f(a_{\ell m}(e)), f(b_{\ell m}(e))) \wedge 2^{\ell m}}{d_{G_k}(e, \widehat{e}_{\ell m}(e)) + 1} \cdot \mathbf{1}_{\{e \frown_{\ell m} \widehat{e}_{\ell m}(e)\}} - \max_{\substack{i \in \{1,\ldots,\ell-1\} \\ e \frown_{im} \widehat{e}_{im}(e)}} \frac{d_T(f(a_{im}(e)), f(b_{im}(e))) \wedge 2^{im}}{d_{G_k}(e, \widehat{e}_{im}(e)) + 1}\right\} = \Omega(\ell). \quad (27)$$

To prove (27) denote for $C \in \mathscr{C}_{[\alpha]}$

$$S_C = \left\{ e \in E(C) :\ \varepsilon_C \frown_{\ell m} e \ \text{ and } \max_{\substack{i \in \{1,\ldots,\ell-1\} \\ e \frown_{im} \widehat{e}_{im}(e)}} \frac{d_T(f(a_{im}(e)), f(b_{im}(e))) \wedge 2^{im}}{d_{G_k}(e, \widehat{e}_{im}(e)) + 1} \geq \frac{1}{2} \cdot \frac{d_T(f(a_{\ell m}(e)), f(b_{\ell m}(e))) \wedge 2^{\ell m}}{d_{G_k}(e, \widehat{e}_{\ell m}(e)) + 1} \right\}.$$

Then using (26) we see that

$$\begin{aligned}
A &\geq \frac{1}{2 \cdot 4^{m\ell}} \sum_{C \in \mathscr{C}_{[\alpha]}} \sum_{\substack{e \in E(C) \setminus S_C \\ \varepsilon_C \frown_{\ell m} e}} \frac{d_T(f(a_{\ell m}(e)), f(b_{\ell m}(e))) \wedge 2^{\ell m}}{d_{G_k}(e, \widehat{e}_{\ell m}(e)) + 1} \\
&\geq \frac{1}{2 \cdot 4^{m\ell}} \sum_{C \in \mathscr{C}_{[\alpha]}} \sum_{\substack{e \in E(C) \\ \varepsilon_C \frown_{\ell m} e}} \frac{d_T(f(a_{\ell m}(e)), f(b_{\ell m}(e))) \wedge 2^{\ell m}}{d_{G_k}(e, \widehat{e}_{\ell m}(e)) + 1} - \frac{1}{2 \cdot 4^{m\ell}} \sum_{C \in \mathscr{C}_{[\alpha]}} \sum_{e \in S_C} \frac{d_T(f(a_{\ell m}(e)), f(b_{\ell m}(e))) \wedge 2^{\ell m}}{d_{G_k}(e, \widehat{e}_{\ell m}(e)) + 1} \\
&\geq \frac{1}{2 \cdot 4^{m\ell}} \sum_{C \in \mathscr{C}_{[\alpha]}} \sum_{i=1}^{2^{m\ell-1}} \frac{2^{m\ell}}{12 i} - \frac{1}{2 \cdot 4^{m\ell}} \sum_{C \in \mathscr{C}_{[\alpha]}} \sum_{e \in S_C} \frac{d_T(f(a_{\ell m}(e)), f(b_{\ell m}(e))) \wedge 2^{\ell m}}{d_{G_k}(e, \widehat{e}_{\ell m}(e)) + 1} \\
&= \Omega\left(\frac{1}{4^{m\ell}} \cdot |\mathscr{C}_{[\alpha]}| \cdot 2^{m\ell} \cdot m\ell\right) - \frac{1}{2 \cdot 4^{m\ell}} \sum_{C \in \mathscr{C}_{[\alpha]}} \sum_{e \in S_C} \frac{d_T(f(a_{\ell m}(e)), f(b_{\ell m}(e))) \wedge 2^{\ell m}}{d_{G_k}(e, \widehat{e}_{\ell m}(e)) + 1} \\
&= \Omega(m\ell) - \frac{1}{2 \cdot 4^{m\ell}} \sum_{C \in \mathscr{C}_{[\alpha]}} \sum_{e \in S_C} \frac{d_T(f(a_{\ell m}(e)), f(b_{\ell m}(e))) \wedge 2^{\ell m}}{d_{G_k}(e, \widehat{e}_{\ell m}(e)) + 1}. \qquad (28)
\end{aligned}$$

To estimate the negative term in (28) fix $C \in \mathscr{C}_{[\alpha]}$. For every edge $e \in S_C$ (which implies in particular that $\widehat{e}_{\ell m}(e) = \varepsilon_C$) we fix an integer $i < \ell$ such that $e \frown_{im} \widehat{e}_{im}(e)$ and

$$\frac{2^{im}}{d_{G_k}(e, \widehat{e}_{im}(e)) + 1} \geq \frac{d_T(f(a_{im}(e)), f(b_{im}(e))) \wedge 2^{im}}{d_{G_k}(e, \widehat{e}_{im}(e)) + 1} \geq \frac{1}{2} \cdot \frac{d_T(f(a_{\ell m}(e)), f(b_{\ell m}(e))) \wedge 2^{\ell m}}{d_{G_k}(e, \widehat{e}_{\ell m}(e)) + 1} \geq \frac{1}{12} \cdot \frac{2^{\ell m}}{d_{G_k}(e, \varepsilon_C) + 1},$$

or

$$d_{G_k}(e, \widehat{e}_{im}(e)) + 1 \leq 2^{(i-\ell)m+4} \left[d_{G_k}(e, \varepsilon_C) + 1 \right]. \qquad (29)$$

We shall call the edge $\widehat{e}_{im}(e)$ the designated edge that inserted e into S_C. For a designated edge $\varepsilon \in E(C)$ of level im (i.e. $\varepsilon \in \Delta_{im}(C)$) we shall denote by $\mathscr{E}_C(\varepsilon)$ the set of edges of C which ε inserted to S_C. Denoting $D_\varepsilon = d_{G_k}(\varepsilon, \varepsilon_C) + 1$ we see that (29) implies that for $e \in \mathscr{E}_C(\varepsilon)$ we have

$$\left|D_\varepsilon - [d_{G_k}(e, \varepsilon_C) + 1]\right| \le 2^{(i-\ell)m+4}\left[d_{G_k}(e, \varepsilon_C) + 1\right]. \tag{30}$$

Assuming that $m \ge 5$ we are assured that $2^{(i-\ell)m+4} \le \frac{1}{2}$. Thus (30) implies that

$$\frac{D_\varepsilon}{1 + 2^{(i-\ell)m+4}} \le d_{G_k}(e, \varepsilon_C) + 1 \le \frac{D_\varepsilon}{1 - 2^{(i-\ell)m+4}}.$$

Hence

$$\begin{aligned}
\sum_{e\in S_C} \frac{d_T(f(a_{\ell m}(e)), f(b_{\ell m}(e))) \wedge 2^{\ell m}}{d_{G_k}(e, \widehat{e}_{\ell m}(e)) + 1} &\le \sum_{i=1}^{\ell-1} \sum_{\varepsilon\in\Delta_{im}(C)} \sum_{e\in\mathscr{E}_C(\varepsilon)} \frac{2^{\ell m}}{d_{G_k}(e, \varepsilon_C) + 1} \\
&\le 2\sum_{i=1}^{\ell-1} \sum_{\varepsilon\in\Delta_{im}(C)} \sum_{\substack{j\in\mathbb{N} \\ \frac{D_\varepsilon}{1+2^{(i-\ell)m+4}} \le j \le \frac{D_\varepsilon}{1-2^{(i-\ell)m+4}}}} \frac{2^{\ell m}}{j} \\
&= O(1)\cdot 2^{\ell m} \sum_{i=1}^{\ell-1} |\Delta_{im}(C)| \cdot \log\left(\frac{1 + 2^{(i-\ell)m+4}}{1 - 2^{(i-\ell)m+4}}\right) \\
&= O(1)\cdot 2^{\ell m}\ell \cdot 2^{(\ell-i)m} \cdot 2^{(i-\ell)m} = O(1)\cdot 2^{\ell m}\ell.
\end{aligned}$$

Thus, using (28) we see that

$$A = \Omega(m\ell) - O(1)\cdot \frac{1}{4^{\ell m}} \cdot \left|\mathscr{C}_{[\alpha]}\right| 2^{m\ell}\ell = \Omega(m\ell) - O(1)\ell = \Omega(\ell),$$

provided that m is a large enough absolute constant.

This completes the proof of the lower bound in Theorem 1. □

4 Monotone clustering problems

In this section we give some examples which illustrate how certain monotone clustering problems can be solved efficiently on ultrametrics. Our arguments are quite flexible, and apply in more general situations. Before passing to these algorithms, we make a few general remarks on the framework for monotone clustering that was discussed in the introduction.

In the definition of monotone clustering we required that $\Gamma(x, d, P)$ is homogeneous in d. One might wonder whether it is possible to consider also higher orders of homogeneity, i.e. clustering cost functions Γ which satisfy $\Gamma(x, \lambda d, P) = \lambda^p \Gamma(x, d, P)$ for some $p > 1$. For the proof of Theorem 2 to work in this setting we need a distribution over non-contractive embeddings into ultrametrics $f : X \to U$ with a polylogarithmic upper bound on the expected value of $|\nabla f(x)|_\infty^p$. Unfortunately, this is impossible to achieve in general. Indeed, let $f : C_n \to T$ be a random non-contractive embedding of the n-cycle into trees. Lemma 6 implies that there exists an edge $(x, x+1) \in E(C_n)$ for which $d_T(f(x), f(x+1)) \ge \frac{n}{3} - 1$. Thus

$$\sum_{\{x,y\}\in E(C_n)} d_T(f(x), f(y))^p \ge \frac{n^p}{12^p}.$$

Taking expectation we see that

$$\max_{x\in V(C_n)} \mathbb{E}|\nabla f(x)|_\infty^p \geq \frac{1}{n}\sum_{x\in V(C_n)} \mathbb{E}|\nabla f(x)|_\infty^p \geq \frac{n^{p-1}}{12^p}.$$

We note, however, that the proof of Theorem 2 used the homogeneity of Γ in a weak way. In order to get a polylogarithmic reduction to ultrametrics is enough to assume, for example, that for every $\lambda \geq 1$ we have $\Gamma(x, \lambda d, P) = O(\text{polylog}(n)) \cdot \lambda \cdot \Gamma(x, d, P)$.

Our second remark concerns the fact that the solution space for monotone clustering problem that was presented in the introduction was $2^{X\times 2^X}$. This is a huge space, and as we have seen in Section 1.1, by setting the clustering cost function to be ∞ on certain possible clustering solutions it is possible to reduce the size of this space. Additionally, in the arguments is Section 1.1 the cost function Γ ignored the structure of the solution space. Thus in a more generic formulation of the monotone clustering framework we can assume that the solution space is some abstract finite set $S(X)$. For example, in our version of the fault-tolerant k-median problem we can take the solution space to be $\binom{X}{k}$.

4.1 Monotone clustering on ultrametrics via dynamic programming

We now pass to the design of some monotone clustering algorithms on ultrametrics. It is a standard fact (see for example [6]) that any ultrametric (U, d_U) can be represented as follows. There is a graph theoretical tree $T = (V, E)$ such that U is the set of leaves of T. The vertices of T are labelled by $\Delta : V \to [0, \infty)$ and for every $u, v \in U$ we have $d_U(u, v) = \Delta(\text{lca}(u, v))$, where $\text{lca}(u, v)$ is the least common ancestor of u and v in T. We may, and will, assume in what follows that every vertex of T is either a leaf or has exactly two children.

We begin by showing that the fault-tolerant version of the k-median problem described in (5) can be solved exactly on ultrametrics.

Lemma 10. *The minimization of the objective function in* (5) *can be solved exactly on any n-point ultrametric in time $O(kn^2)$.*

Proof. Let (U, d_U) be an n-point ultrametric and let $T = (V, E)$ be a binary tree with vertex labels $\Delta : V \to [0, \infty)$ which represents U. We also assume that we are given fault-tolerant parameters $\{j(u)\}_{u\in U}$. For every $v \in V$ let T_v denote the subtree of T rooted at v. Define for $v \in V$ and $s \in \{0, \ldots, k\}$

$$\text{cost}^*(v, s) = \min\left\{ \sum_{\substack{x\in T_v\cap U\\ j(x)\leq s}} d_U\left(x, x^*_{j(x)}(x; d_U)\right) :\ x_1, x_2, \ldots, x_s \in T_v \cap U \right\}. \tag{31}$$

Our goal is to compute $\text{cost}^*(r, k)$, where r is the root of T. This will be done using dynamic programming. For any leaf $u \in U$ and $s \in \{0, \ldots, k\}$ define $\text{cost}(u, s) = 0$. Let $v \in V$ be an internal vertex with two children $u, w \in V$. Define recursively

$$\begin{aligned}\text{cost}(v, s) = \min_{t\in\{0,\ldots,s\}} \Big[&\text{cost}(u, t) + \text{cost}(w, s - t) \\ &+ \Delta(v) \cdot \left(|\{x \in T_u \cap U :\ t < j(x) \leq s\}| + |\{x \in T_w \cap U :\ s - t < j(x) \leq s\}|\right)\Big]. \end{aligned}\tag{32}$$

A bottom-up computation of the dynamic program in (32) computes $\text{cost}(v, s)$ naïvely in $O(kn^2)$ time. We will be done if we show that $\text{cost}(v, s) = \text{cost}^*(v, s)$ for any $v \in V$ and $s \in \{0, \ldots, k\}$. The fact that

$\mathrm{cost}^*(v,s) \le \mathrm{cost}(v,s)$ is obvious since (32) computes a feasible solution of (31) (this fact is proved by a straightforward induction).

We prove the reverse inequality by induction on the distance of v from the leaves of T. Let $x_1,\ldots,x_s \in T_v \cap U$ be such that

$$\mathrm{cost}^*(v,s) = \sum_{\substack{x\in T_v\cap U\\ j(x)\le s}} d_U\left(x, x^*_{j(x)}(x;d_U)\right).$$

Let u,w be the children of v in T. We may reorder the points so that for some $t \in \{0,\ldots,s\}$ we have $\{x_1,\ldots,x_t\} = T_u \cap \{x_1,\ldots,x_s\}$ and $\{x_{t+1},\ldots,x_s\} = T_w \cap \{x_1,\ldots,x_s\}$. Then

$$\begin{aligned}
\mathrm{cost}^*(v,s) &= \sum_{\substack{x\in T_v\cap U\\ j(x)\le s}} d_U\left(x, x^*_{j(x)}(x;d_U)\right)\\
&= \sum_{\substack{x\in T_u\cap U\\ j(x)\le t}} d_U\left(x, x^*_{j(x)}(x;d_U)\right) + \sum_{\substack{x\in T_w\cap U\\ j(x)\le s-t}} d_U\left(x, x^*_{j(x)}(x;d_U)\right)\\
&\quad +\Delta(v)\cdot\left(|\{x\in T_u\cap U:\ t<j(x)\le s\}| + |\{x\in T_w\cap U:\ s-t<j(x)\le s\}|\right) &(33)\\
&\ge \mathrm{cost}^*(u,t) + \mathrm{cost}^*(w,s-t)\\
&\quad +\Delta(v)\cdot\left(|\{x\in T_u\cap U:\ t<j(x)\le s\}| + |\{x\in T_w\cap U:\ s-t<j(x)\le s\}|\right) &(34)\\
&\ge \mathrm{cost}(u,t) + \mathrm{cost}(w,s-t)\\
&\quad +\Delta(v)\cdot\left(|\{x\in T_u\cap U:\ t<j(x)\le s\}| + |\{x\in T_w\cap U:\ s-t<j(x)\le s\}|\right) &(35)\\
&\ge \mathrm{cost}(v,s), &(36)
\end{aligned}$$

where in (33) we used the fact that the tree T represents the ultrametric (U,d_U), in (34) we used the definition of $\mathrm{cost}^*(u,t)$ and $\mathrm{cost}^*(w,s-t)$ given by (31), in (35) we used the inductive hypothesis, and in (36) we used (32). □

Our final result is the proof of Lemma 3, which yields a FPTAS for the $\Sigma\ell_p$ clustering problem on ultrametrics. We start with the following inequality.

Lemma 11. *Fix $p \ge 1$ and assume that $a_1 \ge a_2 \ge \cdots \ge a_n \ge 0$ and $b_1,\ldots,b_n \ge 0$. Then*

$$\sum_{j=1}^n (a_j^p + b_j^p)^{1/p} \ge \sum_{j=2}^n a_j + \left(a_1^p + \sum_{j=1}^n b_j^p\right)^{1/p}.$$

Proof. The proof is by induction on n, and the inductive hypothesis simplifies to

$$\left(a_1^p + \sum_{j=1}^n b_j^p\right)^{1/p} - a_{n+1} \ge \left(a_1^p + \sum_{j=1}^{n+1} b_j^p\right)^{1/p} - (a_{n+1}^p + b_{n+1}^p)^{1/p}. \tag{37}$$

Denote for $x \ge 0$

$$f(x) = \left(a_1^p + \sum_{j=1}^n b_j^p + x\right)^{1/p} - (a_{n+1}^p + x)^{1/p}.$$

Inequality (37) is $f(b_{n+1}^p) \le f(0)$, so it is enough to prove that f is decreasing. But

$$f'(x) = \frac{1}{p\left(a_1^p + \sum_{j=1}^n b_j^p + x\right)^{1-1/p}} - \frac{1}{p\left(a_{n+1}^p + x\right)^{1-1/p}} \le \frac{1}{p\left(a_1^p + x\right)^{1-1/p}} - \frac{1}{p\left(a_{n+1}^p + x\right)^{1-1/p}} \le 0,$$

since $a_1 \ge a_{n+1}$. □

Proof of Lemma 3. Let (U, d_U) be an n-point ultrametric and let $T = (V, E)$ be a binary tree with vertex labels $\Delta : V \to [0, \infty)$ which represents U. For $v \in V$, $\ell \in \{0, \dots, k\}$, $s \in \{0, \dots, n\}$ and $t \in [0, \infty)$ define $B^*(v, \ell, s, t)$ to be the minimum cost according to (7) to cluster $T_v \cap U$ using ℓ sets and centers, when we are allowed to exclude s points from $T_v \cap U$, and the most costly cluster has cost t.

We next define a "pseudo cost" $B(v, \ell, s, t)$ inductively as follows. If v is a leaf then define $B(v, 1, 0, 0) = B(v, 1, 1, 0) = B(v, 0, 1, 0) = 0$, and for all other values of ℓ, s, t we set $B(v, \ell, s, t) = \infty$. When v has children u and w define:

$$B(v,\ell,s,t) = \min\left\{ \begin{array}{l} B(u,\ell_1,s_1,t_1) + B(w,\ell_2,s_2,t_2) \\ \quad + \left(t_1^p + r_2\Delta(v)^p\right)^{1/p} - t_1 + \left(t_2^p + r_1\Delta(v)^p\right)^{1/p} - t_2 : \begin{array}{c} s_1,r_1,s_2,r_2\in\{0,\dots,s\}, \\ t_1,t_2\in[0,t], \\ \ell_1\in\{0,\dots,\ell\}, \\ r_1\le s_1, \\ r_2\le s_2, \\ s=s_1+s_2-r_1-r_2, \\ \ell=\ell_1+\ell_2, \\ t=\max\left\{\left(t_1^p+r_2\Delta(v)^p\right)^{1/p},\ \left(t_2^p+r_1\Delta(v)^p\right)^{1/p}\right\} \end{array} \end{array} \right\}.$$

With these definition we will prove the following claim by induction.

Claim 12. *For every $v \in T$, $\ell \in \{0, \dots, k\}$, $s \in \{0, \dots, n\}$ and $t \in [0, \infty)$ we have*

$$B^*(v, \ell, s, t) = B(v, \ell, s, t).$$

Assuming the validity of Claim 12 for the moment, we conclude as follows. The dynamic programming algorithm described above does not suffice since the parameter t takes values in the range $[0, \infty)$, while we need it to take only $\mathrm{poly}(n)$ values. We fix this issue using an argument which is based on ideas from [5].

Normalize the distances in U so that the minimum distance is 1, and denote $\Phi = \mathrm{diam}(U)$. We can clearly assume that $t \le n\Phi$. Assume first of all that we can ensure that $t \le A = O(\mathrm{poly}(n))$. Once this is achieved then all we need to do is to apply a standard discretization procedure as follows. Fix an integer $M > 0$ which will be determined presently and let $A' = \{0, A/M, 2A/M, \dots, A\}$. For $t \in [0, A]$ denote by $\mathrm{rd}(t)$ the rounding of t to its closest value in A'. We can now define a discretized dynamic programming procedure $B'(v, \ell, s, \tau)$, where v, ℓ, s take the same values as in the definition of $B(v, \ell, s, t)$ and $\tau \in A'$. This is done by defining as before for a leaf $v \in U$ $B(v, 1, 0, 0) = B(v, 1, 1, 0) = B(v, 0, 1, 0) = 0$, and for all other values of ℓ, s, τ setting $B(v, \ell, s, \tau) = \infty$. When v has children u and w define:

$$B'(v,\ell,s,\tau) = \min\left\{ \mathrm{rd}\left(\left(\tau_1^p + r_2\Delta(v)^p\right)^{1/p} - \tau_1 + \left(\tau_2^p + r_1\Delta(v)^p\right)^{1/p} - \tau_2\right)\right.$$

$$\left. + B'(u,\ell_1,s_1,\tau_1) + B'(w,\ell_2,s_2,\tau_2) : \begin{array}{c} s_1,r_1,s_2,r_2\in\{0,\ldots,s\},\\ \tau_1,\tau_2\in A',\\ \ell_1\in\{0,\ldots,\ell\},\\ r_1\le s_1,\\ r_2\le s_2,\\ s=s_1+s_2-r_1-r_2,\\ \ell=\ell_1+\ell_2,\\ \tau=\mathrm{rd}\left(\max\left\{\left(\tau_1^p+r_2\Delta(v)^p\right)^{1/p},\ \left(\tau_2^p+r_1\Delta(v)^p\right)^{1/p}\right\}\right) \end{array}\right\}.$$

It is straightforward to check by induction that for any $v\in V$, $\ell\in\{0,\ldots,k\}$, $s\in\{0,\ldots,n\}$ and $t\in[0,A]$ we have

$$|B(v,\ell,s,t) - B'(v,\ell,s,\mathrm{rd}(t))| \le \frac{4|T_v|}{M}.$$

Since the optimal value of the $\Sigma\ell_p$ clustering problem is at least 1 (excluding trivial cases), as this is the smallest distance in U, B' will yield an approximation algorithm for this problem whose multiplicative error is bounded by $1+O(n/M)$. Taking $M=n/\varepsilon$ for some $\varepsilon\in(0,1)$ we obtain the required PTAS.

We therefore need to argue that we can ensure that $t=O(\mathrm{poly}(n))$. Recall that we can assume that $t\le n\Phi$. Let $P=\{(x_1,C_1),\ldots,(x_k,C_k)\}$ be the (yet unknown) optimal solution of the $\Sigma\ell_p$ clustering problem with k-centers on U. Let h be the maximum length appearing in the solution, i.e. $h=\max_{1\le i\le k}\max_{x\in C_i} d_U(x_i,x)$. Fix $\varepsilon\in(0,1)$ and define two "levels" of the tree T by

$$L=\{v\in V:\ \Delta(v)\le h<\Delta(\mathrm{parent}(v))\},$$

and

$$Q=\left\{v\in V:\ \Delta(v)\le\frac{\varepsilon h}{n^2}<\Delta(\mathrm{parent}(v))\right\}.$$

Let T' be the subtree obtained from T by deleting the subtrees $\{T_v\setminus\{v\}\}_{v\in Q}$, and let U' denote the leaves of T'. Equivalently, U' is obtained from U by contracting all distances smaller that $\varepsilon h/n^2$. It is straightforward to check that $\mathrm{cost}_{U'}(P)\le\mathrm{cost}_U(P)\le(1+\varepsilon)\,\mathrm{cost}_{U'}(P)$.

Note that for every $v\in L$ the aspect ratio (i.e. the ratio of the diameter and the shortest distance) of $T'_v\cap U'$ is at most n^2/ε. So, by the above reasoning (in the case of an a priori polynomial bound on t) we can approximate in polynomial time the value of $B^*(v,\ell,s,t)$ up to a factor $1+O(\varepsilon)$. It remains to "glue" these approximate solutions to a solution of the $\Sigma\ell_p$ clustering problem on T. This is done by a (simpler) dynamic programming argument as follows. Denote by $\widehat{T}$ the subtree of T' whose root is the same as that of T' and whose leaves are L. For $v\in\widehat{T}$ let $C^*(v,\ell)$ be the optimal solution of the $\Sigma\ell_p$ clustering problem on $\widehat{T}_v$ with ℓ centers and assuming that the largest distance appearing in the solution is at most h. We calculate $C^*(v,\ell)$ by dynamic programming: For $v\in L$ define $C(v,\ell)=\min_t B^*(v,\ell,0,t)$, and if v has two children u,w in $\widehat{T}$ then

$$C(v,\ell)=\min\{C(u,\ell_1)+C(w,\ell_2):\ \ell_1\in\{0,\ldots,\ell\},\ \ell_1+\ell_2=\ell\}.$$

A straightforward induction shows that $C^*(v,\ell)=C(v,\ell)$.

The only thing that is left to be explained is how to find the value h. This is done by exhaustive search: We try all the $\binom{n}{2}$ possible values of h, do the above procedure for each of them, and take the minimum of the values that we get.

The proof of Lemma 3 will be complete once we prove Claim 12. We first note that $B^*(v,\ell,s,t) \le B(v,\ell,s,t)$. This is true because $B(\cdot)$ represents a feasible solution of $B^*(\cdot)$. The proof of this fact is by induction. If $u,w \in V$ are the children of v in T then there exist $s_1,s_2,t_1,t_2,r_1,r_2,\ell_1,\ell_2$ such that

$$B(v,\ell,s,t) = B(u,\ell_1,s_1,t_1) + B(w,\ell_2,s_2,t_2) + \left(t_1^p + r_2\Delta(v)^p\right)^{1/p} - t_1 + \left(t_2^p + r_1\Delta(v)^p\right)^{1/p} - t_2,$$

where $s_1,r_1,s_2,r_2 \in \{0,\ldots,s\}$, $t_1,t_2 \in [0,t]$, $\ell_1 \in \{0,\ldots,\ell\}$, $r_1 \le s_1$, $r_2 \le s_2$, $s = s_1+s_2-r_1-r_2$, $\ell = \ell_1+\ell_2$, and $t = \max\left\{\left(t_1^p + r_2\Delta(v)^p\right)^{1/p}, \left(t_2^p + r_1\Delta(v)^p\right)^{1/p}\right\}$. By the inductive hypothesis $B(u,\ell_1,s_1,t_1)$ and $B(w,\ell_2,s_2,t_2)$ correspond to feasible solutions of $B^*(\cdot)$ on $T_u \cap U$ and $T_w \cap U$, respectively. Hence $B(v,\ell,s,t)$ corresponds to the following feasible solution: Take the union of the centers in $T_u \cap U$ and $T_w \cap U$ and retain all the current clusters in $T_u \cap U$ and $T_w \cap U$ as is. Next add arbitrary r_1 unclustered points from $T_u \cap U$ (from the pool of s_1 unclustered points that we are assuming exist in $T_u \cap U$) to the cluster with the most weight in $T_w \cap U$, and similarly add r_2 unclustered points from $T_w \cap U$ to the cluster with the most weight in $T_u \cap U$. This creates the required feasible solution.

We next prove by induction that $B^*(v,\ell,s,t) \ge B(v,\ell,s,t)$. Consider the clustering solution at which $B^*(v,\ell,s,t)$ is attained. It corresponds to s excluded leaves $y_1,\ldots,y_s \in T_v \cap U$, k "centers" $x_1,\ldots,x_\ell \in (T_v \cap U)\setminus\{y_1,\ldots,y_s\}$ and a partition $\{C_1,\ldots,C_\ell\}$ of $(T_v \cap U)\setminus\{y_1,\ldots,y_s\}$ such that

$$B^*(v,\ell,s,t) = \sum_{j=1}^{\ell}\left(\sum_{x\in C_j} d(x,x_j)^p\right)^{1/p}.$$

Moreover, assuming without loss of generality that

$$\sum_{x\in C_1} d(x,x_1)^p = \max_{j\in\{1,\ldots,\ell\}} \sum_{x\in C_j} d(x,x_j)^p$$

the definition of $B^*(v,\ell,s,t)$ guarantees that

$$\left(\sum_{x\in C_1} d(x,x_1)^p\right)^{1/p} = t.$$

By reordering the points we may assume that $x_1,\ldots,x_{\ell_1} \in T_u$ and $x_{\ell_1+1},\ldots,x_{\ell_1+\ell_2}, \in T_w$ (recall that $\ell_1+\ell_2 = \ell$). Denote

$$\left|\left(\bigcup_{j=1}^{\ell_1} C_j\right)\cap T_w\right| = r_2 \quad\text{and}\quad \left|\left(\bigcup_{j=\ell_1+1}^{\ell_1+\ell_2} C_j\right)\cap T_u\right| = r_1.$$

Finally, we may assume that

$$t_1 \stackrel{\text{def}}{=} \sum_{x\in C_1\cap T_u} d(x,x_1)^p = \max_{j\in\{1,\ldots,\ell_1\}} \sum_{x\in C_j\cap T_u} d(x,x_j)^p,$$

and

$$t_2 \stackrel{\text{def}}{=} \sum_{x\in C_{\ell_1+1}\cap T_w} d(x,x_{\ell_1+1})^p = \max_{j\in\{\ell_1+1,\ldots,\ell_1+\ell_2\}} \sum_{x\in C_j\cap T_w} d(x,x_j)^p.$$

Denote

$$A_w = \left(\bigcup_{j=1}^{\ell_1} C_j\right)\cap T_w \quad\text{and}\quad A_u = \left(\bigcup_{j=\ell_1+1}^{\ell_1+\ell_2} C_j\right)\cap T_u.$$

We also write $s_1 = |\{y_1, \dots, y_s\} \cap T_u| + r_1$ and $s_2 = |\{y_1, \dots, y_s\} \cap T_w| + r_2$, so that $s = s_1 + s_2 - r_1 - r_2$.

Note that by definition

$$\sum_{j=1}^{\ell_1} \left(\sum_{x \in C_j \cap T_u} d(x, x_j)^p \right)^{1/p} \geq B^*(u, \ell_1, s_1, t_1), \tag{38}$$

and

$$\sum_{j=\ell_1+1}^{\ell_1+\ell_2} \left(\sum_{x \in C_j \cap T_w} d(x, x_j)^p \right)^{1/p} \geq B^*(w, \ell_2, s_2, t_2). \tag{39}$$

Thus

$$\begin{aligned} B^*(v, \ell, s, t) &= \sum_{j=1}^{\ell_1} \left[\sum_{x \in C_j \cap T_u} d(x, x_j)^p + |C_j \cap A_w| \Delta(v)^p \right]^{1/p} + \sum_{j=\ell_1+1}^{\ell_1+\ell_2} \left[\sum_{x \in C_j \cap T_w} d(x, x_j)^p + |C_j \cap A_u| \Delta(v)^p \right]^{1/p} \\ &\geq B^*(u, \ell_1, s_1, t_1) + B^*(w, \ell_2, s_2, t_2) + \left(t_1^p + r_2 \Delta(v)^p \right)^{1/p} - t_1 + \left(t_2^p + r_1 \Delta(v)^p \right)^{1/p} - t_2 && (40) \\ &\geq B(u, \ell_1, s_1, t_1) + B(w, \ell_2, s_2, t_2) + \left(t_1^p + r_2 \Delta(v)^p \right)^{1/p} - t_1 + \left(t_2^p + r_1 \Delta(v)^p \right)^{1/p} - t_2 && (41) \\ &\geq B(v, \ell, s, t), && (42) \end{aligned}$$

where in (40) we used Lemma 11 together with (38) and (39), in (41) we used the inductive hypothesis, and in (42) we used the definition of $B(\cdot)$. This completes the proof of Lemma 3. □

References

[1] N. Alon, R. Karp, D. Peleg, and D. West, *A graph-theoretic game and its application to the k-server problem*, SIAM J. Comput. **24** (1995), no. 1.

[2] V. Arya, N. Garg, R. Khandekar, A. Meyerson, K. Munagala, and V. Pandit, *Local search heuristics for k-median and facility location problems*, SIAM J. Comput. **33** (2004), no. 3, 544–562.

[3] Y. Bartal, *Probabilistic approximation of metric spaces and its algorithmic applications*, 37th Annual Symposium on Foundations of Computer Science (1996), IEEE Comput. Soc. Press, Los Alamitos, CA, 1996, pp. 184–193.

[4] ———, *On approximating arbitrary metrics by tree metrics*, Proceedings of the 30th Annual ACM Symposium on Theory of Computing, ACM, New York, 1998, pp. 161–168.

[5] Y. Bartal, M. Charikar, and D. Raz, *Approximating min-sum k-clustering in metric spaces*, Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing, ACM, New York, 2001, pp. 11–20.

[6] Y. Bartal, N. Linial, M. Mendel, and A. Naor, *On metric Ramsey-type phenomena*, Ann. of Math. (2) **162** (2005), no. 2, 643–709.

[7] Y. Bartal and M. Mendel, *Multi-embedding of metric spaces*, SIAM J. Comput. **34** (2004), no. 1, 248–259.

[8] S. G. Bobkov and C. Houdré, *Some connections between isoperimetric and Sobolev-type inequalities*, Mem. Amer. Math. Soc. **129** (1997), no. 616, viii+111.

[9] M. R. Bridson and A. Haefliger, *Metric spaces of non-positive curvature*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 319, Springer-Verlag, Berlin, 1999.

[10] B. Brinkman and M. Charikar, *On the impossibility of dimension reduction in l_1*, J. ACM **52** (2005), no. 5, 766–788).

[11] S. Buyalo and V. Schroeder, *Embedding of hyperbolic spaces in the product of trees*, Geom. Dedicata **113** (2005), 75–93.

[12] G. Calinescu, H. Karloff, and Y. Rabani, *Approximation algorithms for the 0-extension problem*, SIAM J. Comput. **34** (2004/05), no. 2, 358–372.

[13] M. Charikar and S. Guha, *Improved combinatorial algorithms for facility location problems*, SIAM J. Comput. **34** (2005), no. 4, 803–824.

[14] M. Charikar, S. Guha, É. Tardos, and D. B. Shmoys, *A constant-factor approximation algorithm for the k-median problem*, J. Comput. System Sci. **65** (2002), no. 1, 129–149.

[15] M. Charikar and R. Panigrahy, *Clustering to minimize the sum of cluster diameters*, J. Comput. Syst. Sci. **68** (2004), no. 2, 417–441.

[16] G. Cornuéjols, G. L. Nemhauser, and L. A. Wolsey, *The uncapacitated facility location problem*, Discrete location theory, Wiley-Intersci. Ser. Discrete Math. Optim., Wiley, New York, 1990, pp. 119–171.

[17] A. N. Dranishnikov, *On hypersphericity of manifolds with finite asymptotic dimension*, Trans. Amer. Math. Soc. **355** (2003), no. 1, 155–167.

[18] M. Elkin, Y. Emek, D. A. Spielman, and S.-H. Teng, *Lower-stretch spanning trees*, Proceedings of the 37th Annual ACM Symposium on Theory of Computing, ACM, New York, 2005, pp. 494–503.

[19] J. Fakcharoenphol, S. Rao, and K. Talwar, *A tight bound on approximating arbitrary metrics by tree metrics*, Proceedings of the 35th annual ACM Symposium on Theory of Computing, 2003, pp. 448–455.

[20] M. Gromov, *Metric structures for Riemannian and non-Riemannian spaces*, Progress in Mathematics, vol. 152, Birkhäuser Boston Inc., Boston, MA, 1999. Based on the 1981 French original; With appendices by M. Katz, P. Pansu and S. Semmes. Translated from the French by Sean Michael Bates.

[21] S. Guha and S. Khuller, *Greedy strikes back: improved facility location algorithms*, Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (San Francisco, CA, 1998), ACM, New York, 1998, pp. 649–657.

[22] A. Gupta, M. Hajiaghayi, and H. Räcke, *Oblivious Network Design*, Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms (Miami, FL, 2006), ACM, New York, 2006, pp. 970–979.

[23] A. Gupta, I. Newman, Y. Rabinovich, and A. Sinclair, *Cuts, trees and l_1-embeddings of graphs*, Combinatorica **24** (2004), no. 2, 233–269.

[24] S. Har-Peled and M. Mendel, *Fast construction of nets in low dimensional metrics, and their applications*, SIAM J. Comput. **35** (2006), no. 5, 1148–1184.

[25] K. Jain, M. Mahdian, E. Markakis, A. Saberi, and V. V. Vazirani, *Greedy facility location algorithms analyzed using dual fitting with factor-revealing LP*, J. ACM **50** (2003), no. 6, 795–824.

[26] K. Jain and V. Vazirani, *Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and Lagrangian relaxation*, J. ACM **48** (2001), no. 2, 274–296.

[27] K. Jain and V. V. Vazirani, *An approximation algorithm for the fault tolerant metric facility location problem*, Algorithmica **38** (2004), no. 3, 433–439. Approximation algorithms.

[28] M. R. Korupolu, C. G. Plaxton, and R. Rajaraman, *Analysis of a local search heuristic for facility location problems*, J. Algorithms **37** (2000), no. 1, 146–188.

[29] R. Krauthgamer, J. R. Lee, M. Mendel, and A. Naor, *Measured descent: A new embedding method for finite metrics*, Geom. Funct. Anal. **15** (2005), no. 4, 839–858.

[30] M. Ledoux, *The concentration of measure phenomenon*, Mathematical Surveys and Monographs, vol. 89, American Mathematical Society, Providence, RI, 2001.

[31] J. R. Lee and A. Naor, *Embedding the diamond graph in L_p and dimension reduction in L_1*, Geom. Funct. Anal. **14** (2004), no. 4, 745–747.

[32] ———, *Extending Lipschitz functions via random metric partitions*, Invent. Math. **160** (2005), no. 1, 59–95.

[33] M. Mendel and A. Naor, *Euclidean quotients of finite metric spaces*, Adv. Math. **189** (2004), no. 2, 451–494.

[34] ———, *Ramsey partitions and proximity data structures*, J. European Math. Soc. **9** (2007), no. 2, 253–275, available at `http://arxiv.org/cs/0511084`.

[35] P. B. Mirchandani and R. L. Francis (eds.), *Discrete location theory*, Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons Inc., New York, 1990. A Wiley-Interscience Publication.

[36] A. Naor, Y. Peres, O. Schramm, and S. Sheffield, *Markov chains in smooth Banach spaces and Gromov hyperbolic metric spaces*, available at `http://arxiv.org/math/0410422`.

[37] I. Newman and Y. Rabinovich, *A lower bound on the distortion of embedding planar metrics into Euclidean space*, Discrete Comput. Geom. **29** (2003), no. 1, 77–81.

[38] Y. Rabinovich and R. Raz, *Lower bounds on the distortion of embedding finite metric spaces in graphs*, Discrete Comput. Geom. **19** (1998), no. 1, 79–94.

[39] D. B. Shmoys, É. Tardos, and K. Aardal, *Approximation algorithms for facility location problems (extended abstract)*, Proceedings of the 29th annual ACM Symposium on Theory of Computing, 1997, pp. 265–274.

[40] C. Swamy and D. B. Shmoys, *Fault-tolerant facility location*, Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms (2003), ACM, New York, 2003, pp. 735–736.

[41] M. Talagrand, *The generic chaining*, Springer Monographs in Mathematics, Springer-Verlag, Berlin, 2005.