Inapproximability of Maximum Diameter Clustering for Few Clusters

Henry Fleischmann[*]   Kyrylo Karlov[†]   Karthik C. S.[‡]   Ashwin Padaki[§]
Stepan Zharkov[¶]

Abstract

In the Max-k-Diameter problem, we are given a set of points in a metric space, and the goal is to partition the input points into k parts such that the maximum pairwise distance between points in the same part of the partition is minimized.

The approximability of the Max-k-Diameter problem was studied in the eighties, culminating in the work of Feder and Greene [STOC'88], wherein they showed it is NP-hard to approximate within a factor better than 2 in the ℓ_1 and ℓ_∞ metrics, and NP-hard to approximate within a factor better than 1.969 in the Euclidean metric. This complements the celebrated 2 factor polynomial time approximation algorithm for the problem in general metrics (Gonzalez [TCS'85]; Hochbaum and Shmoys [JACM'86]).

Over the last couple of decades, there has been increased interest from the algorithmic community to study the approximability of various clustering objectives when the number of clusters is fixed. In this setting, the framework of coresets has yielded PTAS for most popular clustering objectives, including k-means, k-median, k-center, k-minsum, and so on.

In this paper, rather surprisingly, we prove that even when $k = 3$, the Max-k-Diameter problem is NP-hard to approximate within a factor of 1.5 in the ℓ_1-metric (and Hamming metric) and NP-hard to approximate within a factor of 1.304 in the Euclidean metric.

Our main conceptual contribution is the introduction of a novel framework called *cloud systems* which embed hypergraphs into ℓ_p-metric spaces such that the chromatic number of the hypergraph is related to the quality of the Max-k-Diameter clustering of the embedded pointset. Our main technical contributions are the constructions of nontrivial cloud systems in the Euclidean and ℓ_1-metrics using extremal geometric structures.

---

[*]Carnegie Mellon University.

[†]Charles University.

[‡]Rutgers University.

[§]University of Pennsylvania.

[¶]Columbia University.

1   Introduction

The Max-k-Diameter problem is the task of optimizing a classic clustering objective, where given a set of points in a metric space, we are required to partition the points into k parts so as to minimize the maximum pairwise distance between points in the same part of the partition (see Section 3 for a formal definition). This clustering objective was actively studied in the eighties under the lens of approximation. One of the main advantages of Max-k-Diameter over other clustering objectives such as k-center, k-means, and k-median is that Max-k-Diameter is not center-based, and thus one does not have to worry about the quality of centers in the applications of this clustering objective.

By the early 1980s, the Max-k-Diameter problem was known to be NP-complete even in the Euclidean plane (implicit in [FPT81]). Thus, the attention of the community turned towards understanding its approximability in various metric spaces. There were a series of works all providing 2-approximation to the Max-k-Diameter problem in general metrics with improved polynomial runtimes [HS85, HS86, Gon85, FG88]. Hochbaum and Shmoys [HS86] showed that this factor cannot be improved (assuming NP$\neq$P) for general metrics. Thus, the focus shifted to ℓ_p-metrics.

Gonzalez [Gon85] showed that the Max-k-Diameter problem in the Euclidean metric is NP-hard to approximate to a factor of 1.732. This was improved by Feder and Greene [FG88] who showed that the Max-k-Diameter problem in the Euclidean metric is NP-hard to approximate to a factor of 1.969. Moreover, they showed that the Max-k-Diameter problem is NP-hard to approximate within a factor better than 2 in both the ℓ_1 and ℓ_∞ metrics. More importantly, all the results of Feder and Greene hold when the input is in two dimensions.

At first glance, it looks like we have hit a road block in understanding the approximability of the Max-k-Diameter problem. On the one hand, we have completely understood its approximability in the general metrics, ℓ_1, and ℓ_∞ metrics (even in the plane!). On the other hand, the small gap in our knowledge about the Euclidean Max-k-Diameter problem is much like other Euclidean clustering problems (such as Euclidean k-center [FG88, Men88]), which have been open for decades with no real tools of attack to bridge the gap in approximability.

Over the last two-and-a-half decades, a lot of effort has been invested by the algorithmic community to understand the approximability of popular clustering objectives when the number of clusters is *fixed* (in ℓ_p-metrics). In fact, a powerful paradigm has emerged to provide polynomial time approximation schemes (PTAS) for all these problems: *Coresets* [Fel20, CSS21]. In the literature, when k is fixed, a PTAS is already known for k-center [BHI02], k-means [KSS10, Che09], k-median [KSS10, Che09], k-minsum [FKKR03], and other popular clustering objectives.

<center>*Does* Max-k-Diameter *in ℓ_p-metrics admit a* PTAS *when k is fixed?*</center>

We remark here that for the closely related Sum-k-Diameter problem where we are required to partition the input points into k parts so as to minimize the sum of the diameter of parts, there is an exact polynomial time algorithm in general metrics when k is fixed [BS15].

Returning to the Max-k-Diameter problem, an argument by Megiddo [Meg90] proves that Max-3-Diameter is NP-hard to approximate to a factor better than 2 in the ℓ_∞-metric by a reduction from the 3-coloring problem. Thus, we explore the question in the Euclidean metric and the ℓ_1-metric.

<center>*What is the complexity/approximability of* Max-k-Diameter
in the ℓ_1-metric and the ℓ_2-metric when k is fixed?</center>

1.1   Our Results. The main contributions of this paper are strong hardness of approximation results for Max-k-Diameter in the Euclidean metric and the ℓ_1-metric, even when $k = 3$. Our results are surprising, as it is unclear *a priori* why Max-3-Diameter does not admit a PTAS much like the aforementioned other clustering objectives. Note that it has been observed in the literature (for example [Meg90]) that Max-k-Diameter can be reduced to the k-coloring problem; since 2-coloring is in P, we have that Max-k-Diameter is in P when $k = 2$.

First, we present our result for the ℓ_1-metric.

THEOREM 1.1. (FORMALIZED IN THEOREM 5.1) *For every $\varepsilon > 0$ and $k \geq 3$, approximating* Max-k-Diameter *in the ℓ_1-metric (and the Hamming metric) to a factor $1.5 - \varepsilon$ is* NP-*hard.*

To the best of our knowledge, there is no known polynomial time algorithm that achieves a factor less than 2 for even Max-3-Diameter in the ℓ_1-metric. Actually, we prove the above NP-hardness for the Max-3-Diameter problem, and the hardness of approximation continues to hold for the Max-k-Diameter problem when $k > 3$, simply because we can reduce a hard instance of the Max-3-Diameter problem to a hard instance of the Max-k-Diameter problem by adding $k - 3$ many outliers to the input instance.

Additionally, since the above result also holds in the Hamming metric, it implies that the Max-3-Diameter problem in the Euclidean metric is NP-hard to approximate to a factor better than $\sqrt{1.5} \approx 1.224$, using the exact same construction. Interestingly, there is a better than 2-approximation algorithm for the Max-3-Diameter in the ℓ_2-metric. As a straightforward consequence of the PTAS in [BHI02] for the k-center problem, there exists an efficient $(\sqrt{2} + \varepsilon)$-approximation algorithm for the Max-k-Diameter problem in the Euclidean metric (for any $\varepsilon > 0$ and constant $k \in \mathbb{N}$).

That said, we are able to exploit the structure of the Euclidean metric to improve on the $\sqrt{1.5}$-factor NP-hardness and prove the following:

THEOREM 1.2. (FORMALIZED IN THEOREM 6.1) *For every $k \geq 3$, approximating* Max-k-Diameter *in the ℓ_2-metric to a factor 1.304 is* NP-*hard.*

This result is particularly noteworthy, since the Euclidean metric is near-isometrically embeddable into all ℓ_p-metrics, and so hardness within a factor of 1.304 extends to all the ℓ_p-metrics.

The result in Theorem 1.2 is surprising on two counts. First, in the literature we do not know how to exploit Euclidean metric structure to obtain higher hardness of approximation factors for sum clustering objectives (such as k-means, k-median, k-minsum) better than the ones that can be derived from the Hamming metric hard instances [CK19, CKL22]. Even in the max clustering objectives (such as k-center and Max-k-Diameter), *planar geometry* is exploited to obtain strong hardness of approximation results [FG88, Men88], and Theorem 1.2, to the best of our knowledge, is the first time where high-dimensional Euclidean geometry is utilized to present strong hardness of approximation results.[1] Second, whenever there are nontrivial approximation algorithms for a computational problem in a specific metric space, it is notoriously hard to improve hardness of approximation results in that metric space for that problem.

Lastly, it is worth noting that our hardness results continue to hold even when the input pointset (of size n) is restricted to $O(\log n)$ dimensions. In the Euclidean metric, this follows from applying the Johnson-Lindenstrauss lemma [JL84]; in the Hamming metric, it follows from our proof (where we can replace the Hadamard codes used in the proof in this paper by small-biased sets [NN90], consequently bringing the dimension down to $\Theta(\log n)$). Finally, we remark that when the number of dimensions is bounded by a constant, the Max-3-Diameter problem is in P.

The state-of-the-art hardness bounds and approximation algorithms for the Max-k-Diameter problem, for constant $k \geq 3$, is summarized in Table 1.

1.2 Related Works. In this subsection, we mention a few related works to provide better context for our results. We consider the following clustering objectives: Max-k-Diameter, k-means, k-median, k-center, k-minsum, Sum-k-Diameter, and Sum-k-Radii, and also restrict our attention below only to the setting where the number of clusters is fixed. We contrast the clustering objectives mentioned above in two different ways:

---

[1] Even in inapproximability results of other geometric optimization problems such as Euclidean Travelling Salesman Problem [Tre00] or (discrete) Euclidean Steiner Tree problem [FGK24], the structure of high-dimensional Euclidean spaces hinders (more than helps) in proving anything better than what can be inferred from the Hamming metric for those problems.

Metric	NP-Hardness Approximation Factor	Polynomial Time Approximation Factor
ℓ_∞	$\mathbf{2} - \varepsilon$ [Meg90]	$\mathbf{2}$ [Gon85]
ℓ_0/ℓ_1	$\mathbf{1.5} - \varepsilon$ [This Paper]	$\mathbf{2}$ [Gon85]
ℓ_2	$\mathbf{1.304}$ [This Paper]	$\mathbf{1.415}$ [BHI02]

Table 1: State-of-the-art Approximability Results for Max-3-Diameter

Center Based vs. Non-Center Based Objectives. While Max-k-Diameter, Sum-k-Diameter, and k-minsum are not center based objectives, the rest of the aforementioned clustering objectives are all center based objectives. This distinction means that center based objectives even in the general metric can be solved exactly in polynomial time in the *discrete* setting, i.e., when the k centers have to be picked from a set of candidate centers given as input. However, in the *continuous* setting, where the centers can be selected from anywhere in the metric space, this is no longer true, for example see [CKL21] (resp. [Meg90]) wherein they prove NP-hardness of approximation results for the k-means and k-median (resp. k-center) in the continuous setting in the ℓ_∞-metric. That said, even for the k-means, k-median, and k-center problems in ℓ_1 and ℓ_2-metrics, it is possible to show that the problems can be reduced from the continuous setting to the discrete setting with just a $1+\varepsilon$ factor loss in the cost for any $\varepsilon > 0$ [Mat00, CSS21]. On the other hand, the Sum-k-Radii problem even in the continuous setting admits a QPTAS [GKK$^+$10] in general metrics.

Min-Sum vs. Min-Max Objectives. From a combinatorial optimization viewpoint, the clustering objectives Max-k-Diameter and k-center are "min-max" objectives whereas the rest are all "min-sum" objectives. This classification is particularly relevant since hardness of approximation results for the "min-sum" clustering objectives (such as [CK19, CKL21, CKL22]) rely on the PCP theorem [AS98, ALM$^+$98, Din07] and other technically heavy PCP-literature tools such as [Raz98, Kho02], whereas the hardness of approximation results for the "min-max" objectives (such as [FG88, Men88, Meg90]) follow from vanilla NP-hardness of classical problems. This is indeed the case for the results in this paper as well, although we have had to develop sophisticated embedding tools to translate the combinatorial hardness to the geometric Max-k-Diameter problem.

1.3 Organization of the Paper. The organization of the paper is as follows. In Section 2, we provide an overview of the proofs of our main results. In Section 3, we overview some definitions and notations related to the clustering and coloring problems considered throughout the paper. In Section 4, we present the central reduction of the paper using the r-cloud system. In Sections 5 and 6, we present specific constructions of r-cloud systems in the ℓ_1-metric and the ℓ_2-metric, respectively. In Section 7, we outline barriers to proving better hardness of approximation. Finally, in Section 8 we mention a couple of open problem stemming from this work.

2 Proof Overview

Clustering and coloring. There is a natural connection between the geometric problem of Max-k-Diameter and the graph theoretic problem of k-coloring. Namely, there is a simple reduction from Max-k-Diameter to k-coloring where given a pointset P, we can construct a graph G_P where points in P correspond to vertices in G_P and edges are drawn between vertices if the distance between their corresponding points exceeds a threshold $\beta > 0$. Then, G_P is k-colorable if and only if P can be partitioned into k clusters, each of which has diameter at most β.

A simple approach. The above reduction from Max-k-Diameter to k-coloring also motivates a simple approach for reducing k-coloring to the approximation version of Max-k-Diameter. Let G be a graph, and suppose we can embed the vertices of G as a set of points P_G in a metric space, such that the following property holds: if two vertices are adjacent in G, the distance between their corresponding points in P_G is at least $r\beta$ (for some $r > 1$), and if the two vertices are nonadjacent, the distance between their corresponding points in P_G is at most β. Then, if G is k-colorable, there is a clustering of P_G with diameter at most β. However, if G is not k-colorable, then any k-clustering of P_G has diameter at least $r\beta$ (i.e., any partition of P_G into k clusters contains at least one cluster of diameter at least $r\beta$). Therefore, approximating Max-k-Diameter within a factor strictly less than r is at least as hard as determining if G is k-colorable.

We formalize this technique as follows. For a metric space (X, dist), a graph $G = (V, E)$, and an approximation ratio $r > 1$, we say that an r-*embedding* of G is a map $\varphi : V \to X$ such that, for some $\beta > 0$,

1. if $(u, v) \notin E$ then $\mathsf{dist}(\varphi(u), \varphi(v)) \leq \beta$, and

2. if $(u, v) \in E$ then $\mathsf{dist}(\varphi(u), \varphi(v)) \geq r\beta$.

We refer to β as the *short distance* of the r-embedding. The first property guarantees that if G is k-colorable, then the optimal k-clustering diameter of $\varphi(V)$ is at most β. The second property guarantees that if $\varphi(V)$ has a k-clustering with diameter less than $r\beta$, then G is k-colorable.

In particular, if $\mathcal{G}$ is a family of graphs on which k-coloring is NP-hard, and φ is an efficiently computable r-embedding of $\mathcal{G}$, then it is NP-hard to approximate Max-k-Diameter within a factor better than r. We now present a proof-of-concept hardness of approximation result using this r-embedding framework.

Warm up in the ℓ_∞-metric. We describe a simple 2-embedding of general graphs into the ℓ_∞-metric. Let $G = (V, E)$ be a graph with $E = \{e_1, \ldots, e_m\}$. Let $\widetilde{G} = (V, \widetilde{E})$ be a directed graph with $\widetilde{E} = \{\widetilde{e_1}, \ldots, \widetilde{e_m}\}$, where each edge $\widetilde{e_i}$ is an arbitrary orientation of e_i. Define a mapping $\varphi : V \to \mathbb{R}^m$ via $\varphi(u) = (a_{u,1}, \ldots, a_{u,m}) \in \mathbb{R}^m$ given by:

$$a_{u,i} = \begin{cases} 1 & \text{if } \widetilde{e_i} = (u, v) \text{ for some } v \in V \\ -1 & \text{if } \widetilde{e_i} = (v, u) \text{ for some } v \in V \\ 0 & \text{if } u \notin e_i \end{cases}.$$

It can be checked that φ is a 2-embedding in the ℓ_∞-metric. Thus, Max-k-Diameter is hard to approximate within a factor of 2 in the ℓ_∞-metric for any $k \geq 3$, proving that the 2-approximation algorithm is optimal in this setting, due to the NP-hardness of 3-coloring.

We can further extend the above NP-hardness to ℓ_p-metrics in the following way. It is known that 3-coloring remains NP-hard even when restricted to 4-regular graphs [GJS76]. When φ is considered in the ℓ_p-metric, it is an r-embedding of 4-regular graphs for

$$r = \left(\frac{2^p + 6}{8} \right)^{1/p}.$$

Thus, Max-k-Diameter (for $k \geq 3$) is hard to approximate within a factor of $\sqrt{5/4}$ in the ℓ_2-metric. However, in the ℓ_1-metric, this approach does not give any nontrivial hardness, suggesting a need for different techniques.

Trevisan's embedding in the ℓ_1-metric. Trevisan utilized Hadamard codes to prove inapproximability results for geometric problems [Tre00]. Recall that for m a power of 2, a Hadamard code is a set of m elements in $\{0, 1\}^m$ whose pairwise ℓ_1 distances are all $m/2$. We outline a direct application of Trevisan's embedding to obtain a nontrivial hardness of approximation result for Max-3-Diameter in the ℓ_1-metric. In particular, we construct a general 5/4-embedding of graphs which are 4-regular and 4-edge-colorable. It can be shown that 3-coloring is NP-hard on this family of graphs.

On an input graph $G = (V, E)$, let m be a power of 2 that is at least $|V|$. Since G is 4-regular and 4-edge

colorable, we can decompose E as the disjoint union of matchings $M_1, M_2, M_3, M_4 \subseteq E$. Then, we embed each $v \in V$ as the concatenation $\varphi(v) = (a_{v,1}, a_{v,2}, a_{v,3}, a_{v,4}) \in \{0,1\}^{4m}$, where the strings $a_{v,i} \in \{0,1\}^m$ satisfy the following properties for each $i \in [4]$:

1. If $(u,v) \in M_i$, then $a_{u,i} = \overline{a_{v,i}}$ and thus $\left\| a_{u,i} - a_{v,i} \right\|_1 = m$.

2. If $(u,v) \notin M_i$, then $\left\| a_{u,i} - a_{v,i} \right\|_1 = m/2$.

One can construct such an embedding by choosing each $a_{v,i}$ to be either a Hadamard codeword or the complement of a Hadamard codeword. If $(u,v) \in E$, then $\left\| \varphi(u) - \varphi(v) \right\|_1 = \frac{5m}{2}$, and otherwise, $\left\| \varphi(u) - \varphi(v) \right\|_1 = 2m$. Therefore, φ is a 5/4-embedding of G, proving that Max-k-Diameter is hard to approximate within a factor of 5/4 in the ℓ_1-metric for any $k \geq 3$.

As shown by these preliminary results, the r-embedding framework is a straightforward way to prove hardness of approximation results for Max-k-Diameter. However, it turns out that using this framework is restrictive, as it forces a one-to-one correspondence between vertices and points, and it forbids distances in the range $(\beta, r\beta)$. We can relax both of these restrictions using a different reduction from k-coloring to Max-k-Diameter. This novel framework, which we call an *r-cloud system*, allows us to obtain significantly better inapproximability results in the ℓ_1-metric and the ℓ_2-metric.

The r-cloud system. We now explain the intuition behind the r-cloud system for reductions from coloring to Max-k-Diameter. In the previous discussion, we reduced from the problem of k-coloring on graphs. For the r-cloud construction, a more natural problem to consider is panchromatic k-coloring on k-uniform hypergraphs.[2] A simple padding argument shows that this problem is at least as hard as k-coloring on graphs (see Remark 3.2). Next, we explain the definition of r-cloud system as a natural framework for reducing this problem to the Max-k-Diameter problem. The formal definition and proof of the reduction are given in Section 4.

Let $\mathcal{H} = (V, E)$ be a k-uniform hypergraph and (X, dist) be a metric space. The goal is to produce a pointset $P \subset X$ such that panchromatic k-coloring on $\mathcal{H}$ reduces to approximating Max-k-Diameter on P within a factor of r. We start by associating each vertex $v \in V$ with a set $P_v \subset X$, where $P := \bigcup_{v \in V} P_v$ and the sets P_v are pairwise disjoint. Note the contrast to the r-embedding framework, in which each vertex corresponded to only a single point in X. We will also define sets P_e for every edge $e \in E$, which we will motivate later. In the next couple of paragraphs, we will specify two conditions, *proximity* and *spread*, that we would like P to satisfy in order to achieve the completeness and soundness guarantees for the reduction from panchromatic k-coloring to Max-k-Diameter.

For the completeness guarantee of the reduction to hold, we must first show that if $\mathcal{H}$ is panchromatic k-colorable then there is a k-clustering of P with low diameter, say at most some threshold β. Given a k-coloring of $\mathcal{H}$, we create a cluster for each of the k colors, and, for each $v \in V$, we place the entire set P_v in the cluster corresponding to the color of v. Since vertices in $\mathcal{H}$ that are of the same color must be nonadjacent, and we would like every cluster to have diameter at most β, the proximity condition that P must satisfy is naturally defined as follows:

(proximity) For every nonadjacent $v, v' \in V$ we have $\mathsf{diam}(P_v \cup P_{v'}) \leq \beta$.

Note that v and itself are also considered nonadjacent.

Second, for the soundness guarantee of the reduction to hold, we must show that if $\mathcal{H}$ is not panchromatic k-colorable, then every k-clustering of P has large diameter—namely, at least $r\beta$. This can be enforced by showing that any k-clustering of P with diameter strictly less than $r\beta$ corresponds to a panchromatic k-coloring of $\mathcal{H}$. A natural way of extracting a k-coloring of $\mathcal{H}$ from a k-clustering of P is to identify each

---

[2]In panchromatic k-coloring, we are given a k-uniform hypergraph, and the task is to color the vertices such that in each hyperedge, every vertex is assigned a distinct color.

vertex $v \in V$ with a designated point $\rho(v) \in P_v$. Given a k-clustering of P, we create a color for each of the k clusters, and we assign each $v \in V$ the color associated with the cluster of $\rho(v)$. Then, provided that the clustering has diameter less than $r\beta$, we have the following condition: for each edge $e = \{v_1, \ldots, v_k\} \in E$, the points $\rho(v_1), \ldots, \rho(v_k)$ must all be in different clusters. To make the condition easier to verify, we localize it to each edge. In particular, we associate the edge e with a set $P_e \subset P$ that contains the points $\rho(v_1), \ldots, \rho(v_k)$. Then, the spread condition that P must satisfy is naturally defined as follows:

> (spread)　　　For every k-clustering of the set P_e with diameter strictly less than $r\beta$,
> the points $\rho(v_1), \ldots, \rho(v_k)$ must all be in different clusters.

We say P is an r-cloud system of $\mathcal{H}$ if it satisfies the above defined proximity and spread conditions. Importantly, neither condition places a restriction on individual distances within the pointset. Indeed, the final pointset P may contain distances in the range $(\beta, r\beta)$, allowing for more general constructions than the r-embedding framework.

For convenience, we further constrain the sets P_e without losing reasonable generality. First, it is most natural for P_e to be disjoint with all P_v for $v \notin e$. To justify this, recall that whether or not an edge e is properly colored depends only on the colors of its constituent vertices $v \in e$. Equivalently, it is reasonable to assume $P_e \subseteq \bigcup_{v \in e} P_v$. Now, let $\tau(e, v) := P_e \cap P_v$. It follows that P_e can be written as

$$P_e = \bigcup_{v \in e} P_e \cap P_v = \bigcup_{v \in e} \tau(e, v).$$

Similarly, it is reasonable to assume that $P_v \subseteq \bigcup_{e \in E, \, v \in e} P_e$. Thus, P_v can be written as

$$P_v = \bigcup_{\substack{e \in E \\ v \in e}} P_e \cap P_v = \bigcup_{\substack{e \in E \\ v \in e}} \tau(e, v).$$

Then, the entire r-cloud construction is characterized by the sets $\tau(e, v)$ over all $e \in E$ and $v \in e$. In other words, an r-cloud system is simply a pair of mappings (ρ, τ) satisfying the proximity and spread conditions. This is the framing used in Definition 4.2.

Since r-cloud systems can be used to reduce panchromatic k-coloring to approximate Max-k-Diameter, then by noting that the panchromatic k-coloring problem on k-uniform hypergraphs is NP-hard for $k \geq 3$, we have the following result:

THEOREM 2.1. (FORMALIZED IN THEOREM 4.1) *For $k \geq 3$, suppose there is an efficiently computable r-cloud system in a metric space (X, dist) for the family of k-uniform hypergraphs. Then, for any $\varepsilon > 0$, approximating Max-k-Diameter in (X, dist) to a factor $r - \varepsilon$ is NP-hard.*

In Sections 5 and 6, we construct a $3/2$-cloud system in the ℓ_1-metric and a 1.304-cloud system in the ℓ_2-metric, both for 3-uniform hypergraphs. This proves that for $k \geq 3$, Max-k-Diameter is hard to approximate within a factor of $3/2$ in the ℓ_1-metric and 1.304 in the ℓ_2-metric.

Constructing a $3/2$-cloud system in the ℓ_1-metric. Let $\mathcal{H} = (V, E)$ be a 3-uniform hypergraph, and for simplicity let $m := |V|$ be a power of 2. We associate each vertex $v \in V$ with a Hadamard codeword $\mathbf{h}_v \in \{0,1\}^m$, and we define $\rho(v) = (\mathbf{h}_v, \mathbf{h}_v) \in \{0,1\}^{2m}$. For each edge $e = \{x, y, z\}$, the set P_e is then comprised of pairs of $\mathbf{h}_x, \mathbf{h}_y, \mathbf{h}_z$ and their complements. Specifically, $\tau(e, x)$ consists of 5 out of the 9 pairs that can be formed using elements in $\left\{ \mathbf{h}_x, \overline{\mathbf{h}}_y, \overline{\mathbf{h}}_z \right\}$, and $\tau(e, y), \tau(e, z)$ are constructed symmetrically.

With this construction, it can be shown that the proximity condition holds with $\beta = m$. For any vertex $v \in V$, recall that P_v is the union of $\tau(e, v)$ over all $e \in E$ that contain v. This means that the only non-complemented codeword used by any point in P_v is $\mathbf{h}_v$, and also that no point in P_v contains the complement $\overline{\mathbf{h}}_v$. As a result, no two points in P_v can contain complementary codewords. By the distance property of the

Hadamard code, $\mathsf{diam}(P_v) \leq m$. In the proof of Lemma 5.1, we slightly extend this argument to show that also $\mathsf{diam}(P_v \cup P_{v'}) \leq m$ for nonadjacent $v, v' \in V$, which is exactly the proximity condition.

Proving that our construction satisfies the spread condition is more involved. We start by fixing an edge $e = \{x, y, z\}$. Then, we argue through casework that any 3-clustering of $P_{\{x,y,z\}}$ in which $(\mathbf{h}_x, \mathbf{h}_x)$ and $(\mathbf{h}_y, \mathbf{h}_y)$ are clustered together necessarily has diameter at least $3m/2$. Since our construction of $\tau(e, \cdot)$ is symmetric in (x, y, z), the spread property follows.

Constructing a 1.304**-cloud system in the** ℓ_2**-metric.** Let $\mathcal{H} = (V, E)$ be a 3-uniform hypergraph with $m := |V|$. We associate each vertex $v \in V$ with a standard basis vector $\mathbf{e}_v \in \mathbb{R}^m$, and we define $\rho(v) = \mathbf{e}_v$. For each edge $e = \{x, y, z\}$, the set P_e is then constructed on the surface of the unit sphere in the three dimensional subspace of $\mathbb{R}^m$ spanned by $\{\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z\}$. Specifically, consider the region on this sphere where the $\mathbf{e}_x$ component is nonnegative, the $\mathbf{e}_y$ and $\mathbf{e}_z$ components are nonpositive, and all other components are 0. We define $\tau(e, x)$ to be a finite net of points on this region. The sets $\tau(e, y)$ and $\tau(e, z)$ are formed symmetrically. An illustration of these regions is shown below in Figure 1.
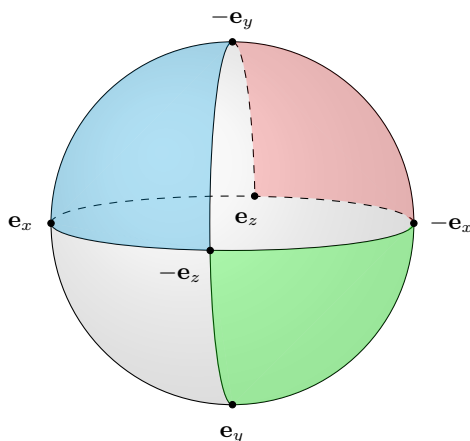

Figure 1: Spherical regions used to define $\tau(e, \cdot)$ for $e = \{x, y, z\}$. Specifically, $\tau(e, x)$, $\tau(e, y)$, and $\tau(e, z)$ are finite nets of the blue, green, and red regions, respectively.

For this construction, the proximity condition holds with $\beta = \sqrt{2}$, and our proof is similar to the proof for the ℓ_1-metric. As before, recall that for any vertex $v \in V$, the set P_v is the union of $\tau(e, v)$ over all $e \in E$ that contain v. This means that the only coordinate on which a point in P_v may be positive is the vth coordinate, and also that no point in P_v has a negative v^{th} component. Since P_v consists of unit vectors, it follows directly that $\mathsf{diam}(P_v) \leq \sqrt{2}$. In the proof of Lemma 6.1, we argue that $\mathsf{diam}(P_v \cup P_{v'}) \leq \sqrt{2}$ for nonadjacent $v, v' \in V$, thus proving the proximity condition.

To prove the spread condition, we fix an edge $e = \{x, y, z\}$. Then, we show that any 3-clustering of $P_{\{x,y,z\}}$ in which $\mathbf{e}_x$ and $\mathbf{e}_y$ are clustered together must have diameter greater than $1.304 \cdot \sqrt{2}$. Since the number of 3-clusterings of P_e is large even for small nets $\tau(e, \cdot)$, we verify this property using a computer search.

We remark that no matter how dense of a net we choose, we cannot improve substantially on the factor 1.304. This is because it is possible to cluster the set $P_{\{x,y,z\}}$ with diameter at most $\sqrt{2 + \sqrt{2}} \approx 1.307 \cdot \sqrt{2}$ such that $\mathbf{e}_x$ and $\mathbf{e}_y$ are in the same cluster. Geometrically, the length $\sqrt{2 + \sqrt{2}}$ is the distance between $\mathbf{e}_y$ and the midpoint of the arc between $\mathbf{e}_x$ and $-\mathbf{e}_y$. Our hardness factor of 1.304 was obtained using a sufficiently dense net for which verifying the spread property was computationally tractable.

Barriers to proving hardness. In Section 7, we outline two barriers to showing improved hardness of approximation for Max-k-Diameter. Recall that one limitation of the r-embedding technique is that it forbids distances strictly between β and $r\beta$ (for some $\beta > 0$). Our first barrier result, formalized in Corollary 7.1, shows that it is not possible to prove hardness above a factor of $5/3$ in the ℓ_1-metric with the r-embedding framework, or using any construction of pointsets with such a gap in the set of distances.

Our next barrier says that if a pointset P in the ℓ_2-metric is not contained within a sphere of diameter $\Delta \cdot \sqrt{2}$, where Δ is the optimal 3-clustering diameter of P, then there is a polynomial time algorithm to approximate Max-3-Diameter on P within a factor of $\sqrt{2} - \varepsilon$. So, in a proof of hardness up to a factor of $\sqrt{2}$ in the ℓ_2-metric, the hard instance must effectively lie within a sphere of diameter $\Delta \cdot \sqrt{2}$. This has an interesting graph theoretic consequence: if Max-3-Diameter is hard to approximate within a factor of 2 in the ℓ_1-metric or within a factor of $\sqrt{2}$ in the ℓ_2-metric, then the graphs associated with the hard instance must have unbounded odd girth. This is formalized in Theorems 7.3 and 7.4.

3 Preliminaries

This section contains definitions, remarks, and notation that will be used throughout this paper.

3.1 Diameter Objective

DEFINITION 3.1. (DIAMETER) *Given a metric space (X, dist) and a set $C \subset X$, we say the diameter of C is*

$$\mathsf{diam}(C) := \max_{x,y \in C} \mathsf{dist}(x, y).$$

For a collection of subsets $C_1, \ldots, C_k \subset X$, we say

$$\mathsf{diam}(\{C_1, \ldots, C_k\}) := \max_{\substack{i \in [k] \\ x,y \in C_i}} \mathsf{dist}(x, y)$$

is the diameter of the collection.

DEFINITION 3.2. (k-CLUSTERING) *Given a set P, we say that a collection $\{C_1, \ldots C_k\}$ is a k-clustering of P if $C_1, \ldots C_k$ are pairwise disjoint and $C_1 \cup \cdots \cup C_k = P$.*

DEFINITION 3.3. (Max-k-Diameter CLUSTERING PROBLEM) *Let (X, dist) be a metric space and let k be a constant. Given as input a finite set $P \subset X$, find a k-clustering $\{C_1, \ldots, C_k\}$ of P that minimizes $\mathsf{diam}(\{C_1, \ldots, C_k\})$.*

DEFINITION 3.4. (r-APPROXIMATE Max-k-Diameter CLUSTERING PROBLEM) *Let (X, dist) be a metric space and let k be a constant. Given as input a finite set $P \subset X$, let*

$$\Delta := \min \mathsf{diam}(\{C_1, \ldots, C_k\}),$$

where the minimum is taken over all possible k-clusterings of P. Find a k-clustering $\{C_1, \ldots, C_k\}$ with diameter at most $r\Delta$.

REMARK 3.1. *Let P be a pointset with optimal k-clustering diameter Δ, and let p be a point whose distance to P is more than $r\Delta$. Then, r-approximate Max-k-Diameter on P reduces to r-approximate Max-$(k+1)$-Diameter on $P \cup \{p\}$, because any $(k+1)$-clustering whose diameter is at most $r\Delta$ must have $\{p\}$ as a standalone cluster. Thus, if Max-k-Diameter clustering is NP-hard, then Max-$(k+1)$-Diameter is NP-hard. In other words, if Max-3-Diameter is hard to approximate to some factor r, then so is Max-k-Diameter for any $k \geq 3$.*

We now restate the definition of r-embedding mentioned in Section 1.

DEFINITION 3.5. (r-EMBEDDING) *Let $G = (V, E)$ be a graph and (X, dist) a metric space. For $r > 1$, we say a map $\varphi : V \to X$ is an r-embedding of G with short distance $\beta > 0$ if the following two conditions hold for every $u, v \in V$:*

- $(u, v) \notin E \implies \mathsf{dist}(\varphi(u), \varphi(v)) \leq \beta$

- $(u, v) \in E \implies \mathsf{dist}(\varphi(u), \varphi(v)) \geq r\beta$

DEFINITION 3.6. *Let k be a constant, $r > 1$, and P a pointset. Let Δ be the optimal k-clustering diameter of P. Define $\Gamma_{k,r}(P)$ to be the graph in which the points in P are vertices, with edges between any two points whose distance is greater than $r\Delta$.*

3.2 Coloring Problems

DEFINITION 3.7. *A hypergraph $\mathcal{H}$ consists of a set of vertices V and a set of hyperedges E, where each hyperedge $e \in E$ is simply a subset $e \subseteq V$. We say that a hypergraph $\mathcal{H}$ is k-uniform if $|e| = k$ for every $e \in E$.*

We define the following coloring problem on k-uniform hypergraphs.

DEFINITION 3.8. (PANCHROMATIC k-COLORING PROBLEM) *Let k be a constant. Given a k-uniform hypergraph $\mathcal{H} = (V, E)$ as input, find an assignment of k colors to the vertices of G such that for every hyperedge $e \in E$, all vertices in e are given distinct colors. Such an assignment is called a panchromatic k-coloring.*

REMARK 3.2. *Note that 3-regular graphs can be viewed as 3-uniform 2-regular hypergraphs: edges correspond to vertices, and triples of edges incident to the same vertex correspond to hyperedges. Since 3-edge coloring is NP-hard on 3-regular graphs [Hol81], we know that panchromatic 3-coloring is NP-hard on 3-uniform hypergraphs.*

3.3 Hadamard Codes. For a bit string $x \in \{0, 1\}^m$, we denote its bitwise complement by $\overline{x}$.

DEFINITION 3.9. (HADAMARD CODE) *Let m be a power of 2. There exists a subset of $\{0, 1\}^m$ of size m such that the Hamming distance between distinct elements is exactly $m/2$. We call this a Hadamard code and denote it by Had_m^+. We also define $\mathrm{Had}_m^- := \{\overline{\mathbf{h}} : \mathbf{h} \in \mathrm{Had}_m^+\}$, which can alternatively be seen as an affine shift of Had_m^+ by the vector $\mathbf{1} = (1, 1, \ldots, 1)$ in the vector space $\mathbb{F}_2^m$. Then, let*

$$\widetilde{\mathrm{Had}}_m := \mathrm{Had}_m^+ \cup \mathrm{Had}_m^-.$$

Given $\mathbf{h}_i, \mathbf{h}_j \in \widetilde{\mathrm{Had}}_m$, we will denote the concatenation of $\mathbf{h}_i$ and $\mathbf{h}_j$ by $(\mathbf{h}_i, \mathbf{h}_j) \in \{0, 1\}^{2m}$.

REMARK 3.3. *The set $\widetilde{\mathrm{Had}}_m$ has the following property: for every $\mathbf{h} \in \widetilde{\mathrm{Had}}_m$, there is exactly one element in $\widetilde{\mathrm{Had}}_m$ that is distance m away from $\mathbf{h}$, namely $\overline{\mathbf{h}}$. All other elements have distance $m/2$ to $\mathbf{h}$.*

4 The r-cloud System

In this section, we introduce a geometric object that can be used to prove inapproximability results for the Max-k-Diameter problem.

DEFINITION 4.1. *For a hypergraph $\mathcal{H} = (V, E)$, let $I_{\mathcal{H}} \subset E \times V$ denote its set of incidences:*

$$I_{\mathcal{H}} = \{(e, v) : e \in E, v \in e\}.$$

DEFINITION 4.2. (r-CLOUD SYSTEM) *Let (X, dist) be a metric space and $U \subseteq X$ be a finite set of points. Let $\mathcal{H} = (V, E)$ be a k-uniform hypergraph. Suppose that (ρ, τ) is a pair of mappings for which $\rho : V \to U$ is injective, and $\tau : I_{\mathcal{H}} \to \mathcal{P}(U)$ satisfies $\rho(v) \in \tau(e, v)$ for all $(e, v) \in I_{\mathcal{H}}$. For $r > 1$, we say that (U, ρ, τ) is an r-cloud system of $\mathcal{H}$ if the following properties hold for some $\beta > 0$:*

- **(Proximity)** *For $v \in V$, define*

$$P_v := \bigcup_{\substack{e \in E \\ (e,v) \in I_{\mathcal{H}}}} \tau(e, v).$$

For any nonadjacent $v, v' \in V$, we have $\mathsf{diam}(P_v \cup P_{v'}) \leq \beta$. Note that we consider v to be nonadjacent to itself.

- **(Spread)** *For $e \in E$, define*

$$P_e := \bigcup_{v \in e} \tau(e, v)$$

Let $e = \{v_1, \ldots, v_k\} \in E$. For any k-clustering of P_e with diameter strictly less than $r\beta$, the points $\rho(v_1), \ldots, \rho(v_k) \in P_e$ are all in different clusters.

We refer to β as the short distance of (U, ρ, τ). Furthermore, we say that an r-cloud system (U, ρ, τ) is efficiently computable if there are $\mathrm{poly}(|V|)$-time algorithms to identify U and evaluate ρ and τ.

The existence of an efficiently computable r-cloud system gives a polynomial time reduction from panchromatic k-coloring to Max-k-Diameter.

THEOREM 4.1. *Let $k \in \mathbb{N}$ be a constant and let $\mathcal{H} = (V, E)$ be a k-uniform hypergraph. Suppose $\mathcal{H}$ has an efficiently computable r-cloud system (U, ρ, τ) with short distance β, where $U \subseteq X$ for a metric space (X, d). Let*

$$P := \bigcup_{(e,v) \in I_{\mathcal{H}}} \tau(e, v) = \bigcup_{v \in V} P_v = \bigcup_{e \in E} P_e.$$

Then, there is a polynomial time algorithm which takes as input $\mathcal{H}$ and outputs P with the following guarantees:

- **Completeness:** *If $\mathcal{H}$ is panchromatic k-colorable, then there exists a k-clustering of P with diameter at most β.*

- **Soundness:** *If $\mathcal{H}$ is not panchromatic k-colorable, then any k-clustering of P has diameter at least $r\beta$.*

Proof. We note that, since $|I_{\mathcal{H}}| \leq n^k$, the pointset P is computable in polynomial time.

We first prove completeness. Let $\mathcal{H}$ be panchromatic k-colorable, and let $c : V \to [k]$ be a panchromatic k-coloring of $\mathcal{H}$. We define a k-clustering $\{C_1, \ldots, C_k\}$ of P given by

$$C_i := \bigcup_{v \,:\, c(v) = i} P_v.$$

To show that this clustering has diameter at most β, it suffices to show that for any $i \in [k]$ and $x, y \in C_i$, we have $d(x, y) \leq \beta$. This follows from the observation that $x \in P_v$ and $y \in P_{v'}$ for some $v, v' \in V$, and since $c(v) = c(v')$, either v, v' are nonadjacent or $v = v'$. Applying the proximity property of Definition 4.2 gives $d(x, y) \leq \mathsf{diam}(P_v \cup P_{v'}) \leq \beta$.

Now, we prove soundness. Suppose $\{C_1, \ldots, C_k\}$ is a k-clustering of P with diameter less than $r\beta$. We construct a k-coloring $c : V \to [k]$ as follows: for each $v \in V$, let $c(v) = i$ if $\rho(v) \in C_i$. Now, let $e = \{v_1, \ldots, v_k\} \in E$ be an arbitrary hyperedge. By the spread property of Definition 4.2, we have that $c(v_1), \ldots, c(v_k)$ are all distinct, and thus c is panchromatic. $\qquad\square$

By Remark 3.2, the above theorem gives a framework of proving hardness of approximating Max-k-Diameter.

5 Hardness of Approximation in the ℓ_1-metric

In this section, we demonstrate the existence of an efficiently computable 3/2-cloud system for any 3-uniform hypergraph in the ℓ_1-metric, hence proving that Max-k-Diameter is NP-hard to approximate within a factor of 3/2 in the ℓ_1-metric.

DEFINITION 5.1. *Let $\mathcal{H}$ be a 3-uniform hypergraph, let $m \geq |V|$ be a power of 2. Identify $V = [|V|] \subseteq [m]$ so that each vertex $v \in V$ corresponds to a distinct codeword $\mathbf{h}_v \in \widetilde{\mathrm{Had}}_m$. Let*

$$U = \widetilde{\mathrm{Had}}_m \times \widetilde{\mathrm{Had}}_m \subset \{0,1\}^{2m}$$

We define an injective map $\rho : V \to U$ given by $\rho(v) = (\mathbf{h}_v, \mathbf{h}_v)$. Moreover, we define $\tau : I_{\mathcal{H}} \to \mathcal{P}(U)$ as follows. For every edge $e = \{x,y,z\}$, choose an arbitrary orientation $\widetilde{e} = (x,y,z)$. Then, let τ be given by:

$$\tau(e,x) = \left\{ (\mathbf{h}_x, \mathbf{h}_x), \left(\mathbf{h}_x, \overline{\mathbf{h}}_z\right), \left(\overline{\mathbf{h}}_y, \mathbf{h}_x\right), \left(\overline{\mathbf{h}}_y, \overline{\mathbf{h}}_z\right), \left(\overline{\mathbf{h}}_z, \overline{\mathbf{h}}_z\right) \right\}$$

$$\tau(e,y) = \left\{ (\mathbf{h}_y, \mathbf{h}_y), \left(\mathbf{h}_y, \overline{\mathbf{h}}_x\right), \left(\overline{\mathbf{h}}_z, \mathbf{h}_y\right), \left(\overline{\mathbf{h}}_z, \overline{\mathbf{h}}_x\right), \left(\overline{\mathbf{h}}_x, \overline{\mathbf{h}}_x\right) \right\}$$

$$\tau(e,z) = \left\{ (\mathbf{h}_z, \mathbf{h}_z), \left(\mathbf{h}_z, \overline{\mathbf{h}}_y\right), \left(\overline{\mathbf{h}}_x, \mathbf{h}_z\right), \left(\overline{\mathbf{h}}_x, \overline{\mathbf{h}}_y\right), \left(\overline{\mathbf{h}}_y, \overline{\mathbf{h}}_y\right) \right\}$$

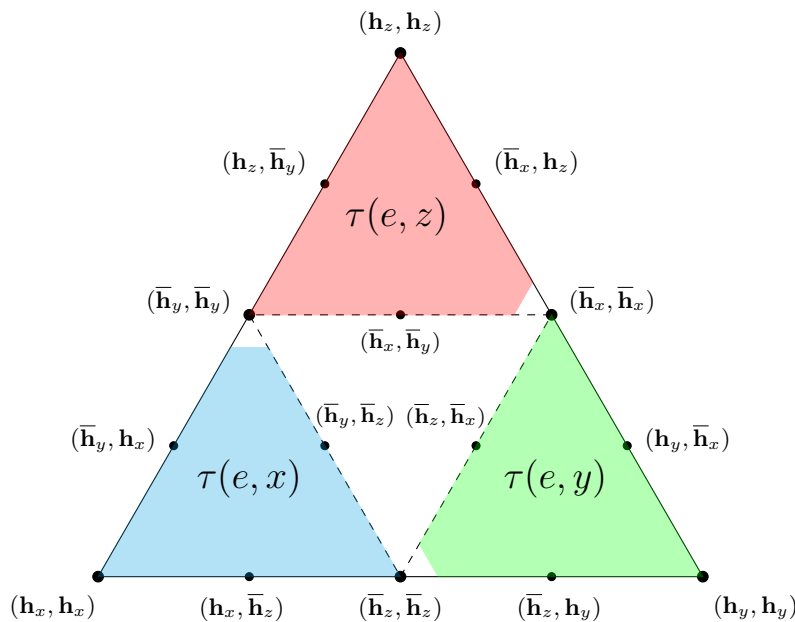*This definition of τ is illustrated in Figure 2.*

Figure 2: Illustration of $\tau(e, \cdot)$ for $\widetilde{e} = (x,y,z)$.

To motivate this choice of τ, we introduce the following terminology. First, for a given point $p = (p_1, p_2) \in U$, we refer to $p_1 \in \widetilde{\mathrm{Had}}_m$ and $p_2 \in \widetilde{\mathrm{Had}}_m$ as the first and second *blocks* of p, respectively. For $A \subseteq U$, we say a Hadamard codeword $\mathbf{h}_v \in \mathrm{Had}_m^+$ *features positively in A* if $\mathbf{h}_v$ is a block (either the first or the second) of some point in A. Similarly, we say $\mathbf{h}_v$ *features negatively in A* if $\overline{\mathbf{h}}_v$ is a block of some point in A.

Our choice of τ, then, has the following *exclusivity* property. For any $e = \{x,y,z\} \in E$, no codeword in Had_m^+ features both positively and negatively in $\tau(e,x)$: the codeword $\mathbf{h}_x$ features positively but not

negatively, while the codewords $\mathbf{h}_y$ and $\mathbf{h}_z$ feature negatively but not positively. From just this property, we can conclude that $\mathsf{diam}(\tau(e,x)) \leq m/2 + m/2 = m$, since no two points in $\tau(e,x)$ may contain complementary first blocks or complementary second blocks. Analogous statements hold for $\tau(e,y)$ and $\tau(e,z)$.

LEMMA 5.1. *Let $\mathcal{H}$ be a 3-uniform hypergraph. Let U, ρ, τ be as in Definition 5.1. Then, (U, ρ, τ) is an efficiently computable 3/2-cloud system of $\mathcal{H}$ with short distance $\beta = m$.*

Proof. It is clear that ρ and τ can each be evaluated in polynomial time. Moreover, since the Hadamard code can be computed in polynomial time, the set U can also be computed in polynomial time.

We will first show that (U, ρ, τ) satisfies the proximity condition in Definition 4.2. Let $v, v' \in V$ be nonadjacent. Then, it suffices to show that for any $p, p' \in P_v \cup P_{v'}$, we have $\left\| p - p' \right\|_1 \leq m$. Writing $p = (p_1, p_2)$ and $p' = (p'_1, p'_2)$ where $p_1, p_2, p'_1, p'_2 \in \widetilde{\mathrm{Had}}_m$, it suffices to show that $\left\| p_1 - p'_1 \right\|_1, \left\| p_2 - p'_2 \right\|_1 \leq m/2$, or equivalently, that $p'_1 \neq \overline{p_1}$ and $p'_2 \neq \overline{p_2}$. It is enough, then, to prove that no codeword in Had_m^+ features both positively and negatively in $P_v \cup P_{v'}$.

Observe that we have

$$
P_v \cup P_{v'} = \left(\bigcup_{\substack{e \in E \\ (e,v) \in I_\mathcal{H}}} \tau(e, v) \right) \cup \left(\bigcup_{\substack{e' \in E \\ (e',v') \in I_\mathcal{H}}} \tau(e', v') \right)
$$

By the exclusivity property of τ, the only codewords featuring positively in $P_v \cup P_{v'}$ are $\mathbf{h}_v$ and $\mathbf{h}_{v'}$. Moreover, $\mathbf{h}_v$ does not feature negatively in P_v, as it does not feature negatively in any $\tau(e, v)$. We claim that $\mathbf{h}_v$ cannot feature negatively in $P_{v'}$ either. Otherwise, there must be some $e' \in E$ containing v' such that $\mathbf{h}_v$ features negatively in $\tau(e', v')$. But then $v \in e'$, contradicting the fact that v, v' are nonadjacent. Therefore, $\mathbf{h}_v$ does not feature negatively in $P_v \cup P_{v'}$. The same reasoning applies to $\mathbf{h}_{v'}$, allowing us to conclude that (U, ρ, τ) satisfies the proximity condition.

Next, we will show that (U, ρ, τ) satisfies the spread condition in Definition 4.2. By symmetry, it suffices to show that any 3-clustering of $P_{\{x,y,z\}}$ in which $(\mathbf{h}_x, \mathbf{h}_x)$ and $(\mathbf{h}_y, \mathbf{h}_y)$ are clustered together necessarily has diameter at least $3m/2$. Let $\{C_1, C_2, C_3\}$ be a 3-clustering of $P_{\{x,y,z\}}$ and suppose for contradiction that $(\mathbf{h}_x, \mathbf{h}_x), (\mathbf{h}_y, \mathbf{h}_y) \in C_1$, where $\mathsf{diam}(C_1) < 3m/2$. Then, no point in C_1 can contain the codewords $\overline{\mathbf{h}}_x$ or $\overline{\mathbf{h}}_y$, and thus

$$
C_1 \subseteq \{(\mathbf{h}_x, \mathbf{h}_x), (\mathbf{h}_y, \mathbf{h}_y), (\mathbf{h}_z, \mathbf{h}_z), (\mathbf{h}_x, \overline{\mathbf{h}}_z), (\overline{\mathbf{h}}_z, \overline{\mathbf{h}}_z), (\overline{\mathbf{h}}_z, \mathbf{h}_y)\}.
$$

We divide into two cases. In the first case, $(\mathbf{h}_z, \mathbf{h}_z) \in C_1$, and so, to ensure $\mathsf{diam}(C_1) < 3m/2$, it must be that

$$
C_1 = \{(\mathbf{h}_x, \mathbf{h}_x), (\mathbf{h}_y, \mathbf{h}_y), (\mathbf{h}_z, \mathbf{h}_z)\}.
$$

Then, $(\overline{\mathbf{h}}_x, \overline{\mathbf{h}}_x), (\overline{\mathbf{h}}_y, \overline{\mathbf{h}}_y), (\overline{\mathbf{h}}_z, \overline{\mathbf{h}}_z) \in C_2 \cup C_3$. Two of these three points must be in the same cluster; without loss of generality, say $(\overline{\mathbf{h}}_x, \overline{\mathbf{h}}_x), (\overline{\mathbf{h}}_y, \overline{\mathbf{h}}_y) \in C_2$. Then, in order to ensure that $\mathsf{diam}(C_2) < 3m/2$, we must have $(\overline{\mathbf{h}}_y, \mathbf{h}_x), (\mathbf{h}_y, \overline{\mathbf{h}}_x) \in C_3$. But then $\mathsf{diam}(C_3) \geq \left\| (\overline{\mathbf{h}}_y, \mathbf{h}_x) - (\mathbf{h}_y, \overline{\mathbf{h}}_x) \right\|_1 = 2m$.

In the second case, $(\mathbf{h}_z, \mathbf{h}_z) \notin C_1$, so we have

$$
C_1 \subseteq \{(\mathbf{h}_x, \mathbf{h}_x), (\mathbf{h}_y, \mathbf{h}_y), (\mathbf{h}_x, \overline{\mathbf{h}}_z), (\overline{\mathbf{h}}_z, \overline{\mathbf{h}}_z), (\overline{\mathbf{h}}_z, \mathbf{h}_y)\}.
$$

We will now show that either C_2 or C_3 necessarily has diameter $\geq 3m/2$. Without loss of generality, $(\mathbf{h}_z, \mathbf{h}_z) \in C_2$, which in turn implies $(\overline{\mathbf{h}}_z, \overline{\mathbf{h}}_x), (\overline{\mathbf{h}}_y, \overline{\mathbf{h}}_z) \in C_3$. But then $(\overline{\mathbf{h}}_y, \mathbf{h}_x), (\mathbf{h}_y, \overline{\mathbf{h}}_x) \in C_2$, a contradiction since $\mathsf{diam}(C_2) \geq \left\| (\overline{\mathbf{h}}_y, \mathbf{h}_x) - (\mathbf{h}_y, \overline{\mathbf{h}}_x) \right\|_1 = 2m$. We conclude that any 3-clustering of $P_{\{x,y,z\}}$ with diameter $< 3m/2$ has $(\mathbf{h}_x, \mathbf{h}_x), (\mathbf{h}_y, \mathbf{h}_y), (\mathbf{h}_z, \mathbf{h}_z)$ in different clusters, as desired. $\square$

This 3/2-cloud system combined with Theorem 4.1 gives us our first main result.

THEOREM 5.1. (Max-k-Diameter IN THE ℓ_1-METRIC) *Let $k \geq 3$ be constant. Given $m \in \mathbb{N}$ and a pointset $P \subset \mathbb{R}^m$ with the ℓ_1-metric, it is* NP-*hard to distinguish between the following two cases:*

- **Completeness:** *There exists a k-clustering of P with diameter at most 1.*

- **Soundness:** *Any k-clustering of P has diameter at least $3/2$.*

Proof. By Lemma 5.1, for any 3-uniform hypergraph $\mathcal{H} = (V, E)$, there is an efficiently computable $3/2$-cloud system (U, ρ, τ) of $\mathcal{H}$ in the ℓ_1-metric, where $U \subset \mathbb{R}^m$ and m is a power of 2. By scaling all points, we can assume that the short distance of the system is 1. Applying Theorem 4.1 and hardness of panchromatic k-coloring on k-uniform hypergraphs, it is NP-hard to distinguish between the two cases for $k = 3$. The statement follows by Remark 3.1. $\quad\square$

Theorem 5.1 also immediately implies an inapproximability result in the ℓ_2-metric.

COROLLARY 5.1. (Max-k-Diameter IN THE ℓ_2-METRIC) *Let $k \geq 3$ be constant. Given $m \in \mathbb{N}$ and a pointset $P \subset \mathbb{R}^m$ with the ℓ_2-metric, it is* NP-*hard to distinguish between the following two cases:*

- **Completeness:** *There exists a k-clustering of P with diameter at most 1.*

- **Soundness:** *Any k-clustering of P has diameter at least $\sqrt{3/2}$.*

Proof. Since the distance between any two points in $\{0, 1\}^m$ in the ℓ_1-metric is the square of the distance between these points in the ℓ_2-metric, any r-cloud system (U, ρ, τ) where $U \subset (\{0, 1\}^m, \ell_1)$ is a $\sqrt{r}$-cloud system when U is treated as a subset of $(\{0, 1\}^m, \ell_2)$. Applying Theorem 4.1 as above, we obtain the desired result. $\quad\square$

6 Hardness of Approximation in the ℓ_2-metric

In this section, we improve upon Corollary 5.1 by treating the ℓ_2-metric directly. Indeed, we demonstrate the existence of a 1.304-cloud system of any 3-uniform hypergraph in the ℓ_2-metric, hence proving hardness of approximation Max-k-Diameter within a factor of 1.304 in this setting.

We first establish the following notation for integer partitions.

DEFINITION 6.1. *For $\kappa, t \in \mathbb{N}$, let $part(\kappa, t) = \{(\alpha_1, \ldots, \alpha_t) \in (\mathbb{N} \cup \{0\})^t : \sum_{i=1}^t \alpha_i = \kappa\}$.*

Now, we define a net of points on particular regions of the surface of a unit sphere.

DEFINITION 6.2. *Let $\mathcal{H} = (V, E)$ be a 3-uniform hypergraph, and identify $V = [m]$ so that each vertex $v \in V$ corresponds to a distinct standard basis vector $\mathbf{e}_v \in \mathbb{R}^m$. For $\kappa \in \mathbb{N}$, define*

$$U_\kappa = \left\{ \frac{\sum_{v \in [m]} \alpha_v \mathbf{e}_v}{\left\| \sum_{v \in [m]} \alpha_v \mathbf{e}_v \right\|_2} : (|\alpha_1|, \ldots, |\alpha_m|) \in part(\kappa, m) \right\}.$$

Geometrically, U_κ is a net of rational points on the surface of the unit sphere in $\mathbb{R}^m$ centered at $\mathbf{0}$. We define an injective map $\rho : V \to U_\kappa$ given by $\rho(v) = \mathbf{e}_v$. Next, we define $\tau_\kappa : I_{\mathcal{H}} \to \mathcal{P}(U)$ as follows. For

every hyperedge $e = \{x, y, z\}$, choose an arbitrary orientation $\widetilde{e} = (x, y, z)$. Then, let τ_κ be given by:

$$\tau_\kappa(e, x) = \left\{ \frac{\alpha_x \mathbf{e}_x - \alpha_y \mathbf{e}_y - \alpha_z \mathbf{e}_z}{\left\| \alpha_x \mathbf{e}_x - \alpha_y \mathbf{e}_y - \alpha_z \mathbf{e}_z \right\|_2} \; : \; (\alpha_x, \alpha_y, \alpha_z) \in part(\kappa, 3) \right\} \setminus \{-\mathbf{e}_y\}$$

$$\tau_\kappa(e, y) = \left\{ \frac{-\alpha_x \mathbf{e}_x + \alpha_y \mathbf{e}_y - \alpha_z \mathbf{e}_z}{\left\| -\alpha_x \mathbf{e}_x + \alpha_y \mathbf{e}_y - \alpha_z \mathbf{e}_z \right\|_2} \; : \; (\alpha_x, \alpha_y, \alpha_z) \in part(\kappa, 3) \right\} \setminus \{-\mathbf{e}_z\}$$

$$\tau_\kappa(e, z) = \left\{ \frac{-\alpha_x \mathbf{e}_x - \alpha_y \mathbf{e}_y + \alpha_z \mathbf{e}_z}{\left\| -\alpha_x \mathbf{e}_x - \alpha_y \mathbf{e}_y + \alpha_z \mathbf{e}_z \right\|_2} \; : \; (\alpha_x, \alpha_y, \alpha_z) \in part(\kappa, 3) \right\} \setminus \{-\mathbf{e}_x\}$$

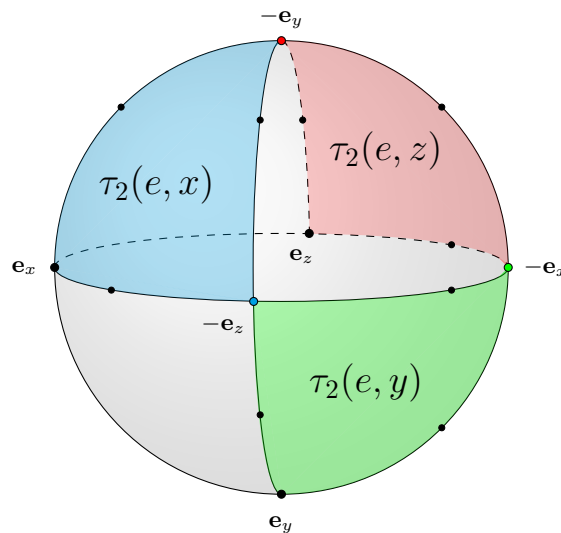*This definition of τ_κ is illustrated in Figure 3.*

Figure 3: Illustration of $\tau_\kappa(e, \cdot)$ for $\widetilde{e} = (x, y, z)$ and $\kappa = 2$. Here, each set $\tau_2(e, \cdot)$ contains 5 points on the surface of a unit sphere. Note the similarity between τ_2 and the construction in the ℓ_1-metric, as shown in Figure 2.

This choice of τ_κ is motivated similarly to the one in Section 5. For $A \subseteq U$, we say that a coordinate $v \in [m]$ *features positively in A* if $p_v > 0$ for some $p \in A$. Similarly, v *features negatively in A* if $p_v < 0$ for some $p \in A$.

Our choice of τ_κ, as in Section 5, has the following *exclusivity* property. For any $e = \{x, y, z\} \in E$, no coordinate in $[m]$ features both positively and negatively in $\tau_\kappa(e, x)$: the coordinate x features positively but not negatively, while the coordinates y and z feature negatively but not positively. Analogous statements hold for $\tau_\kappa(e, y)$ and $\tau_\kappa(e, z)$.

LEMMA 6.1. *Let $U_\kappa, \rho, \tau_\kappa$ be as in Definition 6.2. Then, for a suitable choice of constant κ, $(U_\kappa, \rho, \tau_\kappa)$ is an efficiently computable 1.304-cloud system with short distance $\beta = \sqrt{2}$.*

Proof. For constant $\kappa \in \mathbb{N}$, it is clear that ρ, τ_κ, and U_κ can each be computed in polynomial time.

We will first show that $(U_\kappa, \rho, \tau_\kappa)$ satisfies the proximity condition in Definition 4.2 for any $\kappa \in \mathbb{N}$. Let $v, v' \in V$ be nonadjacent. Then, it suffices to show that for any $p, p' \in P_v \cup P_{v'}$, we have $\left\| p - p' \right\|_2 \leq \sqrt{2}$. Since p and p' are both unit vectors, this is equivalent to proving $\langle p, p' \rangle \geq 0$, as

$$\left\| p - p' \right\|_2 = \sqrt{\langle p - p', p - p' \rangle} = \sqrt{2} - \sqrt{2} \cdot \langle p, p' \rangle.$$

It is enough, then, to show that no coordinate features both positively and negatively in $P_v \cup P_{v'}$, as this implies that $p_i p_i' \geq 0$ for each $i \in [m]$ and thus $\langle p, p' \rangle \geq 0$.

Observe that we have

$$P_v \cup P_{v'} = \left(\bigcup_{\substack{e \in E \\ (e,v) \in I_{\mathcal{H}}}} \tau_\kappa(e,v) \right) \cup \left(\bigcup_{\substack{e' \in E \\ (e',v') \in I_{\mathcal{H}}}} \tau_\kappa(e',v') \right)$$

By the exclusivity property of τ_κ, the only coordinates featuring positively in $P_v \cup P_{v'}$ are v and v'. Moreover, v does not feature negatively in P_v, as it does not feature negatively in any $\tau_\kappa(e,v)$. We claim that v cannot feature negatively in $P_{v'}$ either. Otherwise, there must be some $e' \in E$ containing v' such that v features negatively in $\tau_\kappa(e',v')$. But then $v \in e'$, contradicting the fact that v, v' are nonadjacent. Therefore, v does not feature negatively in $P_v \cup P_{v'}$. The same reasoning applies to v', allowing us to conclude that $(U_\kappa, \rho, \tau_\kappa)$ satisfies the proximity condition for any $\kappa \in \mathbb{N}$.

We claim that $(U_\kappa, \rho, \tau_\kappa)$ satisfies the spread condition in Definition 4.2 for $\kappa = 12$. This can be verified using a computer by confirming that all 3-clusterings of $P_{\{x,y,z\}} \subset U_{12}$ in which $\mathbf{e}_x$ and $\mathbf{e}_y$ are clustered together have diameter at least $1.304 \cdot \sqrt{2}$. The code used to prove the second part of Lemma 6.1 can be found at https://github.com/cea4608937/hardness-of-diameter. $\square$

Applying Theorem 4.1 to this cloud system gives us our main result in the ℓ_2-metric.

THEOREM 6.1. (Max-k-Diameter IN THE ℓ_2-METRIC) *Let $k \geq 3$ be constant. Given $m \in \mathbb{N}$ and a pointset $P \subset \mathbb{R}^m$ with the ℓ_2-metric, it is NP-hard to distinguish between the following two cases:*

- **Completeness:** *There exists a k-clustering of P with diameter at most 1.*

- **Soundness:** *Any k-clustering of P has diameter at least 1.304.*

Proof. By Lemma 6.1, for any 3-uniform hypergraph $\mathcal{H} = (V, E)$, there is an efficiently computable 1.304-cloud system (U, ρ, τ) of $\mathcal{H}$ in the ℓ_2-metric, where $U \subset \mathbb{R}^m$ for $m = |V|$. By scaling all points, we can assume that the short distance of the system is 1. Applying Theorem 4.1 and hardness of panchromatic k-coloring on k-uniform hypergraphs, it is NP-hard to distinguish between the two cases for $k = 3$. The statement follows by Remark 3.1. $\square$

REMARK 6.1. *Fix a hyperedge $e = \{x, y, z\}$, and consider the set $P_{\{x,y,z\}} \subset U_\kappa$. Note that as κ increases, $P_{\{x,y,z\}}$ forms denser nets, allowing us to potentially satisfy the spread condition for larger values of r, thereby obtain larger hardness factors than 1.304. However, there is not much room for improvement with this technique. Namely, the hardness factor we obtain will never surpass $\sqrt{1 + \sqrt{2}/2} \approx 1.307$ because for this particular construction, the spread property does not hold when $r \geq \sqrt{1 + \sqrt{2}/2}$. In particular, we define the following 3-clustering of $P_{\{x,y,z\}} \subset U_\kappa$. Denote the clusters C_x, C_y, C_z, and for $p \in P_{\{x,y,z\}}$, let $p \in C_{v'}$ if*

$$v' = \operatorname*{arg\,min}_{v \in \{x,y,z\}} p_v,$$

deciding ties arbitrarily with a single exception: we force that $\mathbf{e}_x, \mathbf{e}_y \in C_z$, which is possible since $(\mathbf{e}_x)_z = (\mathbf{e}_y)_z = 0$. It can be checked that for any $\kappa \in \mathbb{N}$, the clustering $\{C_x, C_y, C_z\}$ has diameter at most $\sqrt{2 + \sqrt{2}} = \sqrt{2} \cdot \sqrt{1 + \sqrt{2}/2}$. Since $\rho(x)$ and $\rho(y)$ clustered together, this confirms that the spread property does not hold.

7 Barriers to Proving Hardness of Max-k-Diameter

In this section, we present two barriers to showing improved hardness of approximation for the Max-k-Diameter problem.

7.1 Intermediate Distance Barrier. In this section, we show a barrier to our methods of proving hardness of Max-k-Diameter in the ℓ_1-metric. We show that proving hardness of approximation to a factor $r > 5/3$ requires constructing a pointset without large gaps in the set of pairwise distances.

DEFINITION 7.1. *Let $P \subset (X, \mathsf{dist})$ be a set of points. For $a, b \in \mathbb{R}$, we say that P has an (a,b)-gap if there are no $p, p' \in P$ such that $\mathsf{dist}(p, p') \in (a, b)$.*

DEFINITION 7.2. *Let G and H be graphs. We say that G is H-free if no induced subgraph of G is isomorphic to H.*

It turns out that there is a connection between Max-k-Diameter on pointsets with gaps and coloring H-free graphs.

THEOREM 7.1. *Let H be a graph such that H-free k-coloring is in P and H is not r-embeddable. Then, there is a polynomial time algorithm that takes as input pointsets P with a $(\beta, r\beta)$-gap and determines if the optimal 3-clustering of P has diameter at most β.*

Proof. Let $\mathcal{A}$ be a polynomial time algorithm which, given an H-free graph, determines if a k-coloring exists. Given a pointset P with an $(\beta, r\beta)$-gap, construct a graph $G = (V, E)$ where $V = P$ and $(p, p') \in E$ if $\mathsf{dist}(p, p') > \beta$. We observe that the graph G must be H-free, or else the points corresponding to $H \subset G$ would give an r-embedding of H. By running $\mathcal{A}$ on G, we determine if a k-coloring of G exists. A coloring of G directly corresponds to a k-clustering of P with diameter at most β, so this algorithm determines if P has a k-clustering with diameter at most β, as desired. $\square$

In order to apply Theorem 7.1, we must identify a graph H for which H-free k-coloring can be solved in polynomial time.

LEMMA 7.1. *[BCM$^+$18] Let P_7 denote the path graph on 7 vertices. There exists a polynomial time algorithm to 3-color P_7-free graphs.*

COROLLARY 7.1. *For any $r > 5/3$, there is a polynomial time algorithm that takes as input pointsets P in the ℓ_1-metric with a $(\beta, r\beta)$-gap and determines if the optimal 3-clustering of P has diameter at most β.*

Proof. Using the method in Appendix A, we verify that the graph P_7 is not r-embeddable in the ℓ_1-metric for any $r > 5/3$. Applying Lemma 7.1 and Theorem 7.1 gives the result. $\square$

REMARK 7.1. *Corollary 7.1 can be seen as a barrier to various methods of proving hardness of approximation results in the ℓ_1-metric. First, consider the initial r-embedding approach for proving hardness of Max-k-Diameter outlined in Section 1. In this approach, we construct an r-embedding φ (with short distance β) of a hard-to-color graph G, arguing that it is NP-hard to determine if the optimal 3-clustering of $\varphi(G)$ has diameter at most β. Importantly, by definition of r-embedding, the pointset $\varphi(G)$ has a $(\beta, r\beta)$-gap.*

Next, recall the $3/2$-cloud system constructed in Section 5 and the associated pointset P. By the unscaled version of Theorem 5.1, it is NP-hard to determine if the optimal 3-clustering of P has diameter at most m. Moreover, every pairwise distance in $\widetilde{\mathrm{Had}}_m \times \widetilde{\mathrm{Had}}_m$ lies in $\{0, m/2, m, 3m/2, 2m\}$, implying that P has an $(m, 3m/2)$-gap.

Corollary 7.1 implies that any pointset on which Max-k-Diameter cannot be approximated to a $(5/3 + \varepsilon)$ factor must contain intermediate distances, which is not the case in the above two approaches. Note that the approach in Section 6 does have such intermediate distances, so forming nets of points gives one possible method of circumventing this barrier.

7.2 Large Odd Girth Barrier. In this section, we show a better than $\sqrt{2}$-approximation algorithm for Max-3-Diameter, restricted to pointsets that are not contained within any ball of diameter $\Delta \cdot \sqrt{2}$, where Δ denotes the optimal 3-clustering diameter. We then give a graph theoretic implication of this restriction related to the existence of short odd cycles.

We start by formally describing the $(\sqrt{2} + \varepsilon)$-approximation algorithm for Max-k-Diameter mentioned in Section 1. The natural setting of this algorithm is the closely related k-center clustering problem, which we define below.

DEFINITION 7.3. (k-center CLUSTERING PROBLEM) *Let (X, dist) be a metric space and $k \in \mathbb{N}$. Given as input a finite set $P \subset X$, find a k-clustering $\{A_1, \ldots, A_k\} \subset P$ and a collection of "centers" $\{a_1, \ldots, a_k\}$ that minimizes*

$$\max_{i \in [k], \, x \in A_i} \text{dist}(x, a_i).$$

THEOREM 7.2. (FOLLOWS FROM [BHI02]) *Let $\varepsilon > 0$ and k a constant. In Euclidean space, $(\sqrt{2} + \varepsilon)$-approximate k-clustering is in P.*

Proof. From [BHI02], $(1 + \varepsilon)$-approximation of Euclidean k-center is in P. By Jung's theorem [DGK63], any set of points with diameter Δ is contained in a closed ball with radius r satisfying

$$r \leq \Delta \sqrt{\frac{n}{2(n+1)}} \leq \frac{\Delta}{\sqrt{2}}$$

Let Δ be the optimal Max-k-Diameter of the given set of points. Applying Jung's theorem, the optimal clustering for the k-center objective will have diameter

$$\Delta' = 2r \leq \Delta\sqrt{2}$$

Therefore, $(1 + \varepsilon)$-approximation of k-center immediately gives $(\sqrt{2} + \varepsilon)$-approximation of Max-k-Diameter. $\square$

In the case of $k = 3$, we can modify this algorithm to a better approximation for certain pointsets.

THEOREM 7.3. *Let $\varepsilon > 0$ be a sufficiently small constant. Then, there exists a polynomial time $(\sqrt{2} - \varepsilon)$-approximation algorithm for Max-3-Diameter, when restricted to pointsets P not contained in a bounding sphere with diameter $\Delta \cdot \left(\sqrt{2} + 50\varepsilon^{1/8}\right)$, where Δ is the optimal 3-diameter of P.*

The proof of Theorem 7.3 can be found in Appendix B. Importantly, Theorem 7.3 has implications for the graph $\Gamma_{k,r}$ introduced in Definition 3.6.

DEFINITION 7.4. *Given a graph G, the odd girth of G is the length of the smallest odd cycle. If there are no odd cycles, we say that the odd girth is infinite.*

THEOREM 7.4. *For any $g \in \mathbb{N}$, there exists an $\varepsilon > 0$ such that $(\sqrt{2} - \varepsilon)$-approximate Max-3-Diameter is in P, when restricted to pointsets P such that $\Gamma_{k,\sqrt{2}-\varepsilon}(P)$ has odd girth at most g.*

Proof. By scaling, we may assume without loss of generality that the optimal 3-clustering diameter of P is 1. We will prove that for small enough ε, there is no pointset P contained in a ball of diameter $\sqrt{2} + \varepsilon$ such that the odd girth of $\Gamma_{k,\sqrt{2}-\varepsilon}(P)$ is at most g. Then, the statement follows from Theorem 7.3.

So, suppose for contradiction that P is a pointset contained in a ball of diameter $\sqrt{2} + \varepsilon$ and the odd girth of $\Gamma_{k,\sqrt{2}-\varepsilon}(P)$ is at most g. We can assume that this ball is centered at the origin. Since the optimal diameter is 1, all edges in $\Gamma_{k,\sqrt{2}-\varepsilon}(P)$ correspond to distances of at least $\sqrt{2} - \varepsilon$. Let $u, u' \in \Gamma_{k,\sqrt{2}-\varepsilon}(P)$ be arbitrary adjacent vertices and let $p, p' \in P$ be their corresponding points. Note that $\|p - p'\| \geq \sqrt{2} - \varepsilon$.

Since both p and p' are contained in a sphere of diameter $\sqrt{2} + \varepsilon$ centered at the origin, we have that $\left\| p + p' \right\|_2 \leq \frac{1}{2g}$ for small enough ε. Thus, the distance between any two points in P whose corresponding vertices share a neighbor must be at most $2 \cdot \frac{1}{2g} = \frac{1}{g}$ by the triangle inequality.

By assumption, there must be an odd cycle C in $\Gamma_{k,\sqrt{2}-\varepsilon}(P)$ of consisting of at most g vertices. Let $v, v' \in C$ be adjacent vertices and $q, q' \in P$ be their corresponding points. Note that since the length of the cycle C is odd, there is an even length path between v and v' in $\Gamma_{k,\sqrt{2}-\varepsilon}(P)$ that has length at most $g - 1$. Let $q = q_0, q_1, \ldots, q_m = q' \in P$ be every other point along this path, with $m \leq \frac{g}{2}$. For each i, $0 \leq i \leq m-1$, the vertices corresponding to q_i and q_{i+1} share a neighbor, and thus

$$\| q_i - q_{i+1} \| \leq \frac{1}{g}.$$

By the triangle inequality, we have

$$\left\| q - q' \right\|_2 \leq m \cdot \frac{1}{g} \leq \frac{1}{2}.$$

For small enough $\varepsilon > 0$, we have that $1/2 < \sqrt{2} - \varepsilon$, contradicting the fact that q and q' correspond to adjacent vertices. $\square$

REMARK 7.2. *We note that Theorems 7.3 and 7.4 hold only with respect to* Max-3-Diameter*, and not* Max-k-Diameter *for $k \geq 4$. This is in contrast to our inapproximability result from Theorem 6.1, which holds for all $k \geq 3$.*

8   Open Problems

A couple of open problems immediately stem from this work.

- In this work, we proved strong inapproximability results for Max-k-Diameter (where k is fixed) in the Euclidean and ℓ_1-metrics, but our understanding is far from tight in both these metrics, and it remains an important open problem to bridge the gap between the NP-hardness and polynomial time approximability factors. In order to do so, some of the barriers that need to be overcome are detailed in Section 7. However, an intriguing (intermediate) direction that is worth exploring is if we can obtain tight hardness of approximation factors for Max-k-Diameter in the Euclidean metric (or ℓ_1-metric) when k is some large constant.

- We believe that the techniques introduced in this paper, in particular the graph embedding to ℓ_1-metric and Euclidean metric, might be useful to prove improved hardness of approximation results for other geometric problems as well. For example, in [CKL21] the authors gap reduce graph coloring problem to the continuous k-means problem in high-dimensional ℓ_∞-metric, to obtain strong inapproximability results that are higher than the ones derived from covering problems in [CK19, CKL22]. We wonder if it is possible to use our embedding schemes to obtain strong hardness of approximation results for the high dimensional continuous Euclidean k-means problem, improving upon the current state-of-the-art 1.36 NP-hardness factor based on the Johnson Coverage Hypothesis [CKL22].

References

[ALM+98] Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy. Proof verification and the hardness of approximation problems. *J. ACM*, 45(3):501–555, 1998.

[AS98] Sanjeev Arora and Shmuel Safra. Probabilistic checking of proofs: A new characterization of NP. *J. ACM*, 45(1):70–122, 1998.

[BCM+18] Flavia Bonomo, Maria Chudnovsky, Peter Maceli, Oliver Schaudt, Maya Stein, and Mingxian Zhong. Three-coloring and list three-coloring of graphs without induced paths on seven vertices. *Combinatorica*, 38(4), Aug 2018.

[BHI02] Mihai Bādoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 250–257, 2002.

[BS15] Babak Behsaz and Mohammad R Salavatipour. On minimum sum of radii and diameters clustering. *Algorithmica*, 73(1):143–165, 2015.

[Che09] Ke Chen. On coresets for k-median and k-means clustering in metric and Euclidean spaces and their applications. *SIAM Journal on Computing*, 39(3):923–947, 2009.

[CK19] Vincent Cohen-Addad and Karthik C. S. Inapproximability of clustering in l_p-metrics. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science*, pages 519–539, 2019.

[CKL21] Vincent Cohen-Addad, Karthik C. S., and Euiwoong Lee. On approximability of clustering problems without candidate centers. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms SODA*, 2021.

[CKL22] Vincent Cohen-Addad, Karthik C. S., and Euiwoong Lee. Johnson coverage hypothesis: Inapproximability of k-means and k-median in l_p-metrics. In Joseph (Seffi) Naor and Niv Buchbinder, editors, *Proceedings of the 2022 ACM-SIAM Symposium on Discrete Algorithms, SODA 2022, Virtual Conference / Alexandria, VA, USA, January 9 - 12, 2022*, pages 1493–1530. SIAM, 2022.

[CSS21] Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. A new coreset framework for clustering. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 169–182, 2021.

[DGK63] Ludwig Danzer, Branko Grünbaum, and Victor Klee. Helly's theorem and its relatives. In *Proceedings of symposia in pure mathematics*, pages 101–180. American Mathematical Society, 1963.

[Din07] Irit Dinur. The PCP theorem by gap amplification. *Journal of the ACM (JACM)*, 54(3):12–es, 2007.

[Fel20] Dan Feldman. Introduction to core-sets: an updated survey. *arXiv preprint arXiv:2011.09384*, 2020.

[FG88] Tomás Feder and Daniel Greene. Optimal algorithms for approximate clustering. In *Proceedings of the twentieth annual ACM symposium on Theory of computing - STOC '88*, page 434–444, Chicago, Illinois, United States, 1988. ACM Press.

[FGK24] Henry Fleischmann, Surya Teja Gavva, and Karthik C. S. On approximability of Steiner tree in ℓ_p-metrics. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 2024.

[FKKR03] W. Fernandez de la Vega, Marek Karpinski, Claire Kenyon, and Yuval Rabani. Approximation schemes for clustering problems. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 50–58, 2003.

[FPT81] Robert J Fowler, Michael S Paterson, and Steven L Tanimoto. Optimal packing and covering in the plane are NP-complete. *Information processing letters*, 12(3):133–137, 1981.

[GJS76] M. R. Garey, David S. Johnson, and Larry J. Stockmeyer. Some simplified np-complete graph problems. *Theor. Comput. Sci.*, 1(3):237–267, 1976.

[GKK+10] Matt Gibson, Gaurav Kanade, Erik Krohn, Imran A Pirwani, and Kasturi Varadarajan. On metric clustering to minimize the sum of radii. *Algorithmica*, 57:484–498, 2010.

[Gon85] Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.

[Hol81] Ian Holyer. The NP-completeness of edge-coloring. *SIAM Journal on computing*, 10(4):718–720, 1981.

[HS85] Dorit S Hochbaum and David B Shmoys. A best possible heuristic for the k-center problem. *Mathematics of operations research*, 10(2):180–184, 1985.

[HS86] Dorit S. Hochbaum and David B. Shmoys. A unified approach to approximation algorithms for bottleneck problems. *Journal of the ACM*, 33(3):533–550, May 1986.

[JL84] William Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society, 1984.

[Kho02] Subhash Khot. On the power of unique 2-prover 1-round games. In *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada*, pages 767–775, 2002.

[KSS10] Amit Kumar, Yogish Sabharwal, and Sandeep Sen. Linear-time approximation schemes for clustering problems in any dimensions. *Journal of the ACM (JACM)*, 57(2):1–32, 2010.

[Mat00] Jiří Matoušek. On approximate geometric k-clustering. *Discrete & Computational Geometry*, 24(1):61–84, 2000.

[Meg90] Nimrod Megiddo. On the complexity of some geometric problems in unbounded dimension. *Journal of Symbolic Computation*, 10(3–4):327–334, 1990.

[Men88] Stuart G Mentzer. Approximability of metric clustering problems. *Manuscript. https://www. academia. edu/23251714/Approximability of Metric Clustering Problems*, 1988.

[NN90] Joseph Naor and Moni Naor. Small-bias probability spaces: Efficient constructions and applications. In *Proceedings of the twenty-second annual ACM symposium on Theory of computing*, pages 213–223, 1990.

[Raz98] Ran Raz. A parallel repetition theorem. *SIAM J. Comput.*, 27(3):763–803, 1998.

[Tre00] Luca Trevisan. When Hamming meets Euclid: The approximability of geometric TSP and Steiner tree. *SIAM Journal on Computing*, 30(2):475–485, 2000.

A Testing r-embeddability with a Linear Program

It turns out finding the largest ratio r for which a given graph H is r-embeddable into $(\{0,1\}^m, \ell_1)$ can be framed as the solution to a linear program. To see this, let $H = (V, E)$ be a graph on n vertices labeled $v_1, v_2, \ldots, v_n$, and let φ be an r-embedding of H into the space $\{0,1\}^m$, with short distance β. For each $w \in \{0,1\}^n$, associate a nonnegative integer x_w indicating the number of indices $i \in [m]$ such that $(\varphi(v_j))_i = w_j$ holds for all $j \in [n]$. Then, distances in our embedded space can be written as the following linear combination:

$$\big\|\varphi(v_a) - \varphi(v_b)\big\|_0 = \sum_{\substack{w \in \{0,1\}^n \\ w_a \neq w_b}} x_w.$$

Then, for a given short distance parameter β, finding the largest possible embeddability ratio r is equivalent to solving the following optimization problem over nonnegative integer variables $\{x_w : w \in \{0,1\}^n\}$ and a rational variable r:

$$\text{maximize} \quad r$$
$$\text{subject to} \quad \begin{cases} \displaystyle\sum_{\substack{w \in \{0,1\}^n \\ w_a \neq w_b}} x_w \geq r\beta & \text{if } (v_a, v_b) \in E \\ \displaystyle\sum_{\substack{w \in \{0,1\}^n \\ w_a \neq w_b}} x_w \leq \beta & \text{if } (v_a, v_b) \notin E \end{cases}$$

We can remove dependence on s by simply setting $\beta = 1$ and allowing each x_w to take any nonnegative rational value, obtaining the following linear program:

$$\text{maximize} \quad r$$
$$\text{subject to} \quad \begin{cases} \displaystyle\sum_{\substack{w \in \{0,1\}^n \\ w_a \neq w_b}} x_w \geq r & \text{if } (v_a, v_b) \in E \\ \displaystyle\sum_{\substack{w \in \{0,1\}^n \\ w_a \neq w_b}} x_w \leq 1 & \text{if } (v_a, v_b) \notin E \end{cases}$$

Moreover, since each of our constraints have integer coefficients, the solution vector to this linear program will be rational. Thus, we can scale each x_w to be an integer, thus obtaining an embedding into $\{0,1\}^m$ that achieves the optimal ratio r, where $m = \displaystyle\sum_{w \in \{0,1\}^n} x_w \in \mathbb{N}$.

In particular, for a given graph H and parameter $r > 1$, this gives us a computational way to check whether or not H is r-embeddable in the Hamming metric. That said, the number of constraints in the linear program is $\Omega(2^n)$, meaning this method is only tractable when $n = |H|$ is relatively small.

B Proof of Theorem 7.3

Proof. Let $\varepsilon' = 45\varepsilon^{1/4}$ and fix a pointset P not contained in any bounding sphere with diameter $\sqrt{2} + \varepsilon'$. By scaling, we can assume that $\Delta = \frac{\sqrt{2}}{\sqrt{2}+\varepsilon}$ without loss of generality. Let $A \cup B \cup C$ be an optimal 3-clustering of P.

We run the algorithm from Theorem 7.2 on the pointset P, with approximation factor $\sqrt{2} + \varepsilon$, hence obtaining a 3-clustering of P with diameter at most $\Delta \cdot (\sqrt{2} + \varepsilon) = \sqrt{2}$. Theorem 2.7 of [BHI02] gives that all three clusters can be written as the intersection of balls in $\mathbb{R}^m$ with the pointset P, with points contained in multiple balls assigned arbitrarily. Moreover, we desire a stronger property: that the optimal clusters A, B, C are each contained in one of these balls. Slightly extending their algorithm by returning all 3-clusterings enumerated by the simulation of the "guessing oracle", we can ensure that one of the returned 3-clusterings achieves a $(\sqrt{2} + \varepsilon)$-approximation factor as well as the desired property. It suffices to run the following procedure over all of these 3-clusterings. On at least one iteration, the desired approximation will hold.

Let N_A, N_B, N_C be balls of diameter $\sqrt{2}$ that contain A, B, C, respectively. Let a, b, c be their respective centers. We now split into two cases which, by a relabeling argument, are exhaustive.

- **Case I:** $N_A \cap N_B$, $N_A \cap N_C$, $N_B \cap N_C$ each have diameter $\leq \sqrt{2} - \varepsilon$. In this case, we have the following algorithm:

  1. For each point $p \in P$, initialize a set S_p as follows:

  $$S_p = \{X \in \{A, B, C\} : p \in N_X\}$$

  as the set of balls that contain p.

  2. For every pair of points $p, q \in P$ such that $\|p - q\|_2 > \Delta$, if S_p contains only one element, and S_q contains that element, remove it from S_q. Repeat this until no such removals are possible.

  3. Let

  $$A' := \{p \in P : A \in S_p, B \notin S_p\}$$
  $$B' := \{p \in P : B \in S_p, C \notin S_p\}$$
  $$C' := P \setminus (A' \cup B').$$

  Output the 3-clustering $\{A', B', C'\}$.

  We can show that the following property remains true throughout the iterations of step 2 of the algorithm: if p is in some optimal cluster $X \in \{A, B, C\}$, then $X \in S_p$. It is true in the initialization of S_p in step 1 of the algorithm. We only remove X from S_p if there exists some q such that $\|p - q\|_2 > \Delta$, and $q \in X$. So, p cannot be in X, because the optimal clustering $\{A, B, C\}$ has diameter Δ. Thus, if $p \in X$, then $X \in S_p$.

  Let $p, q \in A'$. Note that $A \in S_p$ and $A \in S_q$ by definition. If $S_p = \{A\}$ or $S_q = \{A\}$, then $\|p - q\|_2 \leq \Delta$, or A would have been removed. In all other possible cases, $C \in S_p$ and $C \in S_q$. So, $p, q \in N_A \cap N_C$. Thus, $\|p - q\|_2 \leq \sqrt{2} - \varepsilon$.

  Similarly, if $p, q \in B'$, then either $\|p - q\|_2 \leq \Delta$ or $p, q \in N_A \cap N_B$. If $p, q \in C'$, then either $\|p - q\|_2 \leq \Delta \leq \sqrt{2} - \varepsilon$ or $p, q \in N_C \cap N_B$. In all cases, $\|p - q\|_2 \leq \sqrt{2} - \varepsilon$, so our 3-clustering has diameter at most $\sqrt{2} - \varepsilon$.

- **Case II:** $N_A \cap N_B$ has diameter $> \sqrt{2} - \varepsilon$. We claim that, in this case, the diameter of $N_C \cap (N_A \cup N_B)$ is at most $\sqrt{2} - \varepsilon$. Given that this is true consider the following algorithm:

1. For each point $p \in P$, initialize

$$S_p = \{X \in \{A, B, C\} : p \in N_X\}$$

   as the set of balls that contain p, as in Case I.

2. For every pair of points $p, q \in P$ such that $\|p - q\|_2 > \Delta$, if S_p contains only one element, and S_q contains that element, remove it from S_q. Repeat this until no such removals are possible, as in Case I.

3. Let

$$C' := \{p \in P : C \in S_p\}.$$

   Find an optimal 2-clustering of $P \setminus C'$, and call the output clusters A' and B'. Output the clustering A', B', C'.

Note that for any point $p \in P \setminus C'$, we have that $C \notin S_p$. Thus, $P \setminus C' \subseteq A \cup B$, so there exists a 2-clustering of $P \setminus C'$ of diameter at most Δ. So, A' and B' have diameter at most Δ.

Now, for any $p, q \in C'$, if $S_p = \{C\}$ or $S_q = \{C\}$, then $\|p - q\|_2 \leq \Delta$, or C would have been removed. So, if $\|p - q\|_2 > \Delta$, then both S_p and S_q contain either A or B. In this case, $p, q \in N_C \cap (N_A \cup N_B)$, so $\|p - q\|_2 \leq \sqrt{2} - \varepsilon$. Hence, the clustering $\{A', B', C'\}$ has diameter at most $\sqrt{2} - \varepsilon$.

It remains to verify our claim that $N_C \cap (N_A \cup N_B)$ has diameter at most $\sqrt{2} - \varepsilon$. First, observe that $N_A \cup N_B$ are contained within a ball N_D whose center d is the midpoint between a and b and whose radius is $\frac{\alpha + \sqrt{2}}{2}$, where $\alpha := \|a - b\|_2$. Since $N_A \cap N_B$ has diameter at least $\sqrt{2} - \varepsilon$, by the Pythagorean Theorem we have:

$$\alpha \leq 2\sqrt{\left(\frac{\sqrt{2}}{2}\right)^2 - \left(\frac{\sqrt{2} - \varepsilon}{2}\right)^2} = \sqrt{2\varepsilon\sqrt{2} - \varepsilon^2}.$$

Now, we apply the fact that P is not contained in a bounding sphere of diameter $\Delta \cdot (\sqrt{2} + \varepsilon')$. If $\varepsilon' \geq 2\varepsilon$, we observe that

$$\Delta \cdot (\sqrt{2} + \varepsilon') = \sqrt{2} \cdot \frac{\sqrt{2} + \varepsilon'}{\sqrt{2} + \varepsilon} = \sqrt{2} + \sqrt{2} \cdot \frac{\varepsilon' - \varepsilon}{\sqrt{2} + \varepsilon} \geq \sqrt{2} + \sqrt{2} \cdot \frac{\varepsilon'/2}{\sqrt{2}} = \sqrt{2} + \frac{\varepsilon'}{2}.$$

Hence, it is also true that P is not contained in a bounding sphere of diameter $\sqrt{2} + \frac{\varepsilon'}{2}$. The same is true for $N_C \cup N_D$, since $P \subset N_C \cup N_D$, meaning:

$$\sqrt{2} + \frac{\varepsilon'}{2} \leq \text{diameter}(N_C \cup N_D)$$

$$\leq \|c - d\|_2 + \text{radius}(N_C) + \text{radius}(N_D)$$

$$= \|c - d\|_2 + \frac{\sqrt{2}}{2} + \frac{\alpha + \sqrt{2}}{2}$$

$$= \|c - d\|_2 + \frac{\alpha}{2} + \sqrt{2}.$$

Hence, $\|c - d\|_2 \geq \varepsilon' - \frac{\alpha}{2}$. Finally, let h denote the diameter of $N_C \cap N_D$; since $A \cup B \subset N_D$ and

$C \subseteq N_C$, it suffices to show that $h \leq \sqrt{2} - \varepsilon$. By the Pythagorean Theorem, we have:

$$\frac{h}{2} \leq \sqrt{\text{radius}(N_D)^2 - \left(\frac{\|c-d\|_2}{2}\right)^2}$$

$$\leq \frac{1}{2} \cdot \sqrt{(\alpha + \sqrt{2})^2 - \left(\frac{\varepsilon' - \alpha}{2}\right)^2}$$

$$= \frac{1}{2} \cdot \sqrt{\frac{3\alpha^2}{4} + 2\alpha\sqrt{2} + \frac{\alpha\varepsilon'}{2} + 2 - \frac{\varepsilon'^2}{4}}$$

Hence, it suffices to show that:

$$(\sqrt{2} - \varepsilon)^2 = 2 - 2\varepsilon\sqrt{2} + \varepsilon^2 \geq \frac{3\alpha^2}{4} + \left(2\sqrt{2} + \frac{\varepsilon'}{2}\right) \cdot \alpha + 2 - \frac{\varepsilon'^2}{4}$$

From here, we provide a sequence of stronger inequalities, eventually showing that our setting of $\varepsilon' = 60\varepsilon^{1/4}$ is satisfies the above inequality. First, recall that $\alpha \leq \sqrt{2\varepsilon\sqrt{2} - \varepsilon^2}$. Thus, it is sufficient to have

$$2 - 2\varepsilon\sqrt{2} + \varepsilon^2 \geq \frac{6\varepsilon\sqrt{2} - 3\varepsilon^2}{4} + \left(2\sqrt{2} + \frac{\varepsilon'}{2}\right)\sqrt{2\varepsilon\sqrt{2} - \varepsilon^2} + 2 - \frac{\varepsilon'^2}{4}$$

Subtracting $(2 + \frac{6\varepsilon\sqrt{2} - 3\varepsilon^2}{4})$ from both sides, it is equivalent that

$$\frac{7}{4}(\varepsilon^2 - 2\varepsilon\sqrt{2}) \geq \left(2\sqrt{2} + \frac{\varepsilon'}{4}\right)\sqrt{2\varepsilon\sqrt{2} - \varepsilon^2} - \frac{\varepsilon'^2}{4}$$

Applying the trivial bound $\varepsilon^2 \geq 0$, it is sufficient that

$$-\frac{7\varepsilon}{\sqrt{2}} \geq \left(2\sqrt{2} + \frac{\varepsilon'}{2}\right)\sqrt{2\varepsilon\sqrt{2}} - \frac{\varepsilon'^2}{4}$$

For $\varepsilon' \leq 1$, we have $\left(2\sqrt{2} + \frac{\varepsilon'}{2}\right)\sqrt{2\sqrt{2}} < 6$, so it would be sufficient if

$$-\frac{7\varepsilon}{\sqrt{2}} \geq 6\sqrt{\varepsilon} - \frac{\varepsilon'^2}{4}.$$

Rearranging, we have:

$$\varepsilon'^2 \geq 24\sqrt{\varepsilon} + \frac{28\varepsilon}{\sqrt{2}}.$$

Finally, for $\varepsilon \leq 1$ it is true that $\varepsilon \leq \sqrt{\varepsilon}$, and since $24 + \frac{28}{\sqrt{2}} < 45$, it is good enough to have

$$\varepsilon'^2 \geq 45\sqrt{\varepsilon}$$

This is satisfied when $\varepsilon' = 45\varepsilon^{1/4}$. Note that our argument relied on $\varepsilon \leq 1$, $\varepsilon' \leq 1$ and $\varepsilon' \geq 2\varepsilon$ all holding. One can verify that for $\varepsilon' = 45\varepsilon^{1/4}$, all of these bounds are true when ε is sufficiently small.

In both cases, then, we produced an algorithm which outputs a 3-clustering with diameter at most $\sqrt{2} - \varepsilon$. Since the optimal 3-clustering has diameter $\Delta = \frac{\sqrt{2}}{\sqrt{2}+\varepsilon}$, we achieve an approximation ratio of:

$$\frac{\sqrt{2} - \varepsilon}{\Delta} = \frac{(\sqrt{2} - \varepsilon)(\sqrt{2} + \varepsilon)}{\sqrt{2}} = \frac{2 - \varepsilon^2}{\sqrt{2}} = \sqrt{2} - \frac{\varepsilon^2}{\sqrt{2}}.$$

We have demonstrated a $(\sqrt{2} - \frac{\varepsilon^2}{\sqrt{2}})$-approximation algorithm for pointsets not contained in a bounding sphere of diameter $\Delta \cdot \left(\sqrt{2} + 45\varepsilon^{1/4}\right)$. By a change of variables $\varepsilon \mapsto 2^{1/4}\sqrt{\varepsilon}$, this gives a $(\sqrt{2} - \varepsilon)$-approximation for pointsets not contained in a bounding sphere with diameter

$$\Delta \cdot \left(\sqrt{2} + 45\left(2^{1/4}\sqrt{\varepsilon}\right)^{1/4}\right) = \Delta \cdot \left(\sqrt{2} + 45 \cdot 2^{1/16}\varepsilon^{1/8}\right).$$

Since $45 \cdot 2^{1/16} < 50$, this algorithm also holds for pointsets not contained in a bounding sphere with diameter $\Delta \cdot \left(\sqrt{2} + 50\varepsilon^{1/8}\right)$, which is what we wanted to show. $\square$