Inapproximability of Clustering in Lp-metrics

Vincent Cohen-Addad, Karthik Srikanta

Inapproximability of Clustering in ℓ_p-metrics

Vincent Cohen-Addad*
Sorbonne Université, UPMC Univ Paris 06, CNRS, LIP6, Paris, France
vcohenad@gmail.com

Karthik C. S.†
Weizmann Institute of Science
karthik.srikanta@weizmann.ac.il

Abstract

Proving hardness of approximation for min-sum objectives is an infamous challenge. For classic problems such as the Traveling Salesman problem, the Steiner tree problem, or the k-means and k-median problems, the best known inapproximability bounds for ℓ_p-metrics of dimension $O(\log n)$ remain well below 1.01.

In this paper, we take a significant step to improve the hardness of approximation of the k-means problem in various ℓ_p-metrics, and more particularly on ℓ_1, ℓ_2, Hamming and ℓ_∞ metrics of dimension $\Omega(\log n)$.

We show that it is hard to approximate the k-means objective in $O(\log n)$-dimensional space:

(1) To a factor of 3.94 in the ℓ_∞-metric when centers have to be chosen from a discrete set of locations (i.e., the discrete case). This improves upon the result of Guruswami and Indyk (SODA'03) who proved hardness of approximation for a factor less than 1.01.

(2) To a factor of 1.56 in the ℓ_1-metric and to a factor of 1.17 in the ℓ_2-metric, both in the discrete case. This improves upon the result of Trevisan (SICOMP'00) who proved hardness of approximation for a factor less than 1.01 in both the metrics.

(3) To a factor of 1.07 in the ℓ_2-metric, when centers can be placed at arbitrary locations, (i.e., the continuous case). This improves on a result of Lee-Schmidt-Wright (IPL'17) who proved hardness of approximation for a factor of 1.0013.

We also obtain similar improvements over the state-of-the-art hardness of approximation results for the k-median objective in various ℓ_p-metrics.

Our hardness result given in (1) above, is under the standard NP $\neq$ P assumption, whereas all the remaining results given above are under the *Unique Games Conjecture* (UGC). We can remove our reliance on UGC and prove standard NP-hardness for the above problems but for smaller approximation factors.

Finally, we note that in order to obtain our result for the ℓ_1 and ℓ_∞-metrics in $O(\log n)$-dimensional space we introduce an embedding technique which combines the transcripts of certain communication protocols with the geometric realization of certain graphs.

*Ce projet a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du Programme Appel à projets générique JCJC 2018 portant la référence suivante : ANR-18-CE40-0004-01.

†This work was supported by ERC-CoG grant 772839, the Israel Science Foundation (grant number 552/16), and from the Len Blavatnik and the Blavatnik Family foundation.

1 Introduction

Clustering is a classic, routinely-used process, to solve a large variety of problems. Case in point, it is used to carryout unsupervised learning, or to analyze large amount of data, or to solve information retrieval problems, or to detect communities in social networks. Given a dataset and a metric defined over the data elements, a clustering is a partition of the data such that similar data elements are in the same part. Hence, clustering allows to extract information from the data by identifying data elements that share common features. Clustering problems have thus become of fundamental importance and have received a considerable amount of attention through the years.

Arguably, the k-means and k-median objectives are the most successful models for clustering problems. They have been studied since the sixties and the most popular algorithms used in practice, such as the famous KMEANS++ algorithm [AV07] or Lloyd's method, are designed so as to optimize the classic k-means objective: given a set of points P in a metric space, find a set of k points, called centers, in the metric space so as to minimize the sum of squared distances from each point to its closest center (see Section 1.1 for a slightly more formal definition). Similarly, the k-median objective asks to minimize the sum of distances from each point to its closest center.

Therefore, the question of designing algorithms for optimizing the k-means and k-median objectives has taken a preponderant role in both theory and practice. From a theory perspective, both problems are unfortunately NP-Hard, even when the underlying metric space is the Euclidean plane [MS84] but admits a PTAS when d is a fixed constant [CAKM16, FRS16, KR07, Coh18, CFS18]. Nonetheless, for several applications arising in machine learning and data analysis, the dimension of the point set corresponds to the number of *features* of the datasets, which is large for many datasets. Thus, researchers have considered the k-median and k-means in Euclidean space of arbitrary dimension, and also in more general metric spaces, or through the lenses of parameterized complexity. For general metric spaces, both problems are known to be hard to approximate within respectively a $1 + 2/e \approx 1.73$ and $1 + 8/e \approx 3.94$ factor since the late 90's [GK99]. After a long line of work, the best known approximation algorithms for general metric spaces achieve 2.67 and 9 approximation factors for k-median and k-means respectively [BPR+15, ANSW16]. Then it becomes natural and of practical significance to ask whether there exist better approximation algorithms for the Euclidean metric.

However, our understanding of the Euclidean clustering inputs is pretty limited. The best known hardness of approximation for $O(\log n)$-dimensional ℓ_p metrics, where p is finite, is due to the celebrated result of Trevisan [Tre00] and remains below 1.01. While this result was later extended to ℓ_∞ metrics by Guruswami and Indyk [GI03] who obtained comparable hardness bounds, little progress has been made over the last 15 years and the hardness of approximation for these problems remains below 1.01, even for the ℓ_∞ metric[1]. In fact, showing hardness of approximation in Euclidean space for min-sum objectives is a fundamental challenge. For most of the classic optimization problems in metric spaces such as the traveling salesman problem (TSP), Steiner tree (ST), or k-median

[1]Note that ℓ_∞-metric of large dimensions are not of high interest since hardness of approximation for general metric space directly implies hardness of approximation for the same factor in ℓ_∞-metric of high dimension by applying the Fréchet embedding.

and k-means, the best known hardness of approximation are also obtained through the work of Trevisan [Tre00] and Guruswami and Indyk [GI03], and also remain below 1.01. Even after the advances on hardness of approximation based on the unique games conjecture, no better hardness has been established for these problems in ℓ_p-metrics.

This stands in sharp contrast to the best known approximation ratios for the k-means and k-median problems and creates a somewhat frustrating situation: for example, the current approximation ratio for the k-median in any ℓ_p metric is the same as that for a general metric space (i.e., we are not able to leverage the geometry/topology of the ℓ_p-metric space, for any p), while the hardness of approximation in the ℓ_p-metrics is below 1.01 which is in contrast to the $1 + 2/e$ inapproximability for general metric spaces. A somewhat less frustrating case is the k-means problem, for which Ahmadian et al. [ANSW16] have recently shown how to use the structure of the ℓ_2-metric to obtain an approximation ratio of 6.47 improving upon the approximation ratio of 9 for general metric space. Yet, the best known hardness of approximation for the k-means problem in the ℓ_2-metric remains below 1.01.

This also stands in contrast with the problem of showing NP-hardness for computing exact solutions to clustering problems in ℓ_p-metrics, a question which is much better understood. For example, when parameterized by the number of centers, k, both the k-median and k-means problems are known to be W[1]-Hard, even in $\mathbb{R}^4$, and to not admit a better than $n^{o(k)}$ exact algorithm assuming the Exponential Time Hypothesis [CADMRR18]. However, and perhaps surprisingly, a $(1+\varepsilon)$-approximation algorithm running in time $2^{k/\varepsilon^{O(1)}}nd$ [dlVKKR03, KSS04] is known for the ℓ_2-metric of arbitrary dimension d, while there is no better than $1 + 2/e$-factor approximation (for k-median) and $1 + 8/e$-factor approximation (for k-means) algorithms, for general metrics, running in time $f(k, \varepsilon)n^{o(k)}$ for any arbitrary computable function f, assuming the Gap Exponential Time Hypothesis [CAGK+19].

Yet, in terms of hardness of approximation, no significant progress has been made. Bridging the gap between upper and lower bound on the approximability of the k-means and k-median problems in Euclidean instances is thus an important open problem.

"Discrete" vs "Continuous". Unfortunately, our poor understanding of the k-means and k-median problems does not stop here. To explain this we need to distinguish between two variants of the k-median and k-means problems: the *discrete* and the *continuous*. In the discrete case, centers have to be chosen from a specific set of so-called *candidate centers* that is part of the input, while in the continuous case, centers can be chosen arbitrarily in the ℓ_p-metric space.

While hardness of approximation for the discrete variant has been known since the work of Trevisan [Tre00] and Guruswami and Indyk [GI03] as mentioned earlier, the hardness of the continuous version had remained an open problem for a while. Dasgupta first showed that the problem is NP-Hard in large dimensions [Das08]. A recent work of Awasthi et al. [ACKS15] showed the APX-Hardness of the k-means problem in the Euclidean metric and the inapproximability bound was recently improved to 1.0013 by Lee et al. [LSW17]. Yet, we do not know of a better approximation algorithm for the continuous version and so the best known approximation algorithm achieves a 6.47-

approximation.

Previous Approaches. One of the main roadblock for obtaining higher hardness of approximation is perhaps the "degree constraint". More concretely, the embedding technique that is used by Trevisan [Tre00] for ℓ_p-metrics, where p is finite, and by Guruswami and Indyk [GI03] for the ℓ_∞-metric requires to reduce from a "bounded degree" instance of a covering problem, such as vertex cover on bounded degree graphs. However, the hardness of approximation for these problems is very close to 1 and, combined with the loss induced by the embedding, this cannot lead to a hardness greater than 1.01. For example, the recent approach of Awasthi et al. [ACKS15] and Lee et al. [LSW17] is a reduction from vertex cover on triangle-free graphs which introduces a direct embedding for the k-means problem. Unfortunately, the gap of the reduction is also a function of the degree of the input graph, and so requires that the instance of vertex cover has bounded degree.

We bypass the above barriers in two ways. We first provide better reductions, based on the vertex coverage problem (maximization variant of the vertex cover problem), which through a careful analysis leads to a higher gap in $O(n)$ dimensions. While these reductions are satisfactory for the ℓ_2-metric, since they imply hardness of approximation for the problems in $O(\log n)$-dimensional space using the Johnson-Lindenstrauss lemma [JL84], they only lead to hardness of approximation for the problems in ℓ_1- and $\ell_\infty-$metrics of dimension $\Omega(n)$. We then use a interesting, and perhaps surprising, blend of communication protocol and embedding techniques to extend the result to $O(\log n)$-dimensional space. We discuss these ideas further in Section 1.2.

1.1 Our Results

Given two sets of points P and C in a metric space, we define the k-means cost of P for C as the $\sum_{p\in P}\left(\min_{c\in C}\left(\mathrm{dist}(p,c)\right)^2\right)$ and the k-median cost as the $\sum_{p\in P}\left(\min_{c\in C}\mathrm{dist}(p,c)\right)$. Given a set of points P, the k-means (respectively k-median) objective is the minimum over all C of cardinality k of the k-means (respectively k-median) cost of P for C. Given a point $p\in P$, the contribution to the k-means (respectively k-median) cost of p is $\min_{c\in C}\left(\mathrm{dist}(p,c)\right)^2$ (respectively $\min_{c\in C}\mathrm{dist}(p,c)$).

For every $p\in\mathbb{R}_{\geq 1}\cup\{\infty\}$, we define two quantities $\zeta_1(p)$ and $\zeta_2(p)$ (see Section 3 for details). The behavior of these quantities is quite intricate, but for our purpose it suffices to know their values for $p=1,2$, and ∞. We show that $\zeta_1(1)=1.1416$, $\zeta_2(1)=1.5664$, $\zeta_1(2)\approx 1.06$, $\zeta_2(2)\approx 1.1709$, and as $p\to\infty$, we have $\zeta_1(p)\to\zeta_1(\infty)=\zeta_1(1)$ and $\zeta_2(p)\to\zeta_2(\infty)=\zeta_2(1)$.

1.1.1 Inapproximability Results with Candidate Centers

We start by presenting our results on the "discrete" k-median and k-means problems. In these versions, the centers must be chosen from a specific set of points of the metric. We

start with an informal statement below and note that some of the results are under the *unique games conjecture* (UGC).

Theorem 1.1 (Informal statement)**.** *Given n points and poly(n) candidate centers in $O(\log n)$-dimensional space it is* NP*-hard to approximate*

- *the k-*means *objective within a 1.17 factor in ℓ_2-metric (under* UGC*), 1.56 factor in ℓ_1-metric (under* UGC*), and $1 + 8/e \approx 3.94$ in ℓ_∞-metric.*
- *the k-*median *objective within a 1.06 factor in ℓ_2-metric (under* UGC*), 1.14 factor in ℓ_1-metric (under* UGC*), and $1 + 2/e \approx 1.73$ in ℓ_∞-metric.*

Our above results generalizes to any ℓ_p-metric, with a bound which depends on the underlying metric and on the problem.

Theorem 1.2 (Informal statement of Theorems 7.1 and 7.2)**.** *Let $p \in \mathbb{R}_{\geq 1} \cup \{\infty\}$. Assuming* UGC*, given n points and poly(n) candidate centers in $O(\log n)$ dimensional ℓ_p-metric space it is* NP*-hard to distinguish between the following two cases:*

- **Completeness***: The k-*means *objective (resp. k-*median *objective) is at most β (resp. β'),*
- **Soundness***: The k-*means *objective (resp. k-*median *objective) is at least $\zeta_2(p) \cdot \beta$ (resp. $\zeta_1(p) \cdot \beta'$),*

where β (resp. β') is some positive real number depending only on n.

Note that the hardness of approximation factor for the ℓ_∞-metric, given in Theorem 1.2 is worse than the one stated in Theorem 1.1. This is because, the result in Theorem 1.2 follows from combining the hardness of approximation of the vertex coverage problem (under UGC) with certain graph embeddings into ℓ_p-metrics, whereas, the result in Theorem 1.1 for the ℓ_∞-metric follows from combining the hardness of approximation of the *hypergraph vertex coverage* problem with certain hypergraph embeddings, which currently yield meaningful results only for the ℓ_∞-metric.

1.1.2 Inapproximability Results without Candidate Centers

We then move to the "continuous" version of the problems. Here, centers can be placed anywhere in the metric space.

Theorem 1.3 (k-means in Euclidean metric; Informal statement of Theorem 7.4)**.** *Assuming* UGC*, given n points in $O(\log n)$ dimensional Euclidean space it is* NP*-hard to distinguish between the following two cases:*

- **Completeness***: The k-*means *objective is at most β,*
- **Soundness***: The k-*means *objective is at least $1.07 \cdot \beta$,*

where β is some positive real number depending only on n. Moreover, the above hardness holds even when the n points have all their coordinate entries in $\{0, 1\}$.

For the k-median problem without candidate centers, it is in fact more natural to consider the ℓ_1-metric. Indeed, given a set of points in the ℓ_2-metric, computing the median of this set of points is hard, even in the Euclidean plane and no exact algorithm is known. However, in the case of the ℓ_1-metric, it follows easily, it is as simple as computing the location of the mean of a set of points in the ℓ_2-metric. We thus show the following:

Theorem 1.4 (k-median in ℓ_1-metric; Informal statement of Theorem 7.3)**.** *Assuming* UGC*, given n points in $O(\log n)$ dimensional ℓ_1-metric space it is* NP*-hard to distinguish between the following two cases:*

- **Completeness***: The k-median objective is at most β,*
- **Soundness***: The k-median objective is at least $1.07 \cdot \beta$,*

where β is some positive real number depending only on n. Moreover, the above hardness holds even when the n points have all their coordinate entries in $\{0, 1\}$.

The above theorem is obtained through an intermediate hardness of approximation proof for k-median in the Hamming metric (see Theorem 5.2). Also, Theorem 1.3 can be extended to the Hamming metric so as to obtain a slightly higher inapproximability gap of 1.21 (see Theorem 5.1). Typically, it is possible to extend hardness in the Hamming metric to hardness in the edit metric for similarity search type problems. We formalize this intuition and extend Theorem 1.2 to the edit metric as well (see Theorems B.2 and B.3).

Next, we discuss about the hardness results that we can obtain under the more standard NP $\neq$ P assumption. From the exciting progress on the unique games conjecture [KMS17, DKK+18b, DKK+18a, BKS19, KMS18, BK19], we can get the following *unconditional* NP-hardness for approximate vertex coverage problem.

Theorem 1.5 (Essentially combining [BK19] and [AS19])**.** *There is some $\varepsilon > 0$ and $d_0 \in \mathbb{N}$, such that for all $d_{\min} > d_0$, deciding an instance (G, k) of $(0.9807 - \varepsilon)$-vertex coverage problem on minimum degree $d_{\min}$ graphs is* NP*-hard.*

Now we may define for every $p \in \mathbb{R}_{\geq 1} \cup \{\infty\}$, $\zeta_1'(p)$ and $\zeta_2'(p)$. Again for our purpose we only define them for $p = 1, 2$, and ∞: $\zeta_1(1) = 1.04$, $\zeta_2(1) = 1.15$, $\zeta_1(2) \approx 1.02$, $\zeta_2(2) \approx 1.05$, and as $p \to \infty$, we have $\zeta_1(p) \to \zeta_1(\infty) = \zeta_1(1)$ and $\zeta_2(p) \to \zeta_2(\infty) = \zeta_2(1)$.

We can then combine Theorem 1.5 with the embedding given in the proof of Theorem 1.2 to get the hardness of approximation results of Theorem 1.2 where $\zeta_i(p)$ is now replaced by $\zeta_i'(p)$ for $i \in \{1, 2\}$, and we are no longer reliant on UGC. Similarly, we also get NP-hardness (without UGC) as in Theorems 1.3 and 1.4, but for approximation factors roughly equal to 1.02.

Finally, we summarize in Table 1, the state-of-the-art inapproximability factors for the discrete and continuous cases of the k-means and k-median problems in various metric spaces.

Metric \ Problem	Discrete k-means	Discrete k-median	Continuous k-means	Continuous k-median
General	$1+\frac{8}{e} \approx 3.94$ [GK99]	$1+\frac{2}{e} \approx 1.73$ [GK99]	Not Determined†	Not Determined†
ℓ_0	1.56^{*}	1.14^{*}	1.21^{*}	1.07^{*}
ℓ_1	1.56^{*}	1.14^{*}	Not Determined	1.07^{*}
ℓ_2	1.17^{*}	1.06^{*}	1.07^{*}	Not Determined
ℓ_∞	$1+\frac{8}{e} \approx 3.94$	$1+\frac{2}{e} \approx 1.73$	Not Determined	Not Determined

Table 1: In this table we summarize the state-of-the-art inapproximability for k-means and k-median clustering objectives in various metric spaces for both the discrete and continuous versions of the problem. If a citation is not provided for an entry in the table then it implies that the result was obtained in this paper. Also, hardness of approximation factors obtained under the stronger assumption of the unique games conjecture are star marked. Finally, the two entries which are dagger marked, i.e., the inapproximability for k-means and k-median for the general metric in the continuous case, is not explicitly determined in literature, but it *might* be possible to extend Feige's hard instances of the max-coverage problem [Fei98, GK99] to obtain $1+\frac{8}{e}$ and $1+\frac{2}{e}$ for k-means and k-median respectively for the general metric in the continuous case as well.

1.2 Proof Overview

We now give an overview of our techniques.

1.2.1 Warm up

To better understand the state of the art for the hardness of approximation for clustering problems and the different barriers to obtain hardness of approximation for k-median and k-means, let us recall the result of Guha and Khuller [GK99] who showed the 1+2/e hardness of approximation for k-median in general metric spaces.

Given an instance of the set cover problem, where $\mathcal{S}$ denote the sets and $\mathcal{U}$ denote the universe, create an instance of the k-median (or k-means) instance by creating a candidate center c_S for each set $S \in \mathcal{S}$ and a point p_u to be clustered for each element $u \in \mathcal{U}$. Then, set the distances from p_u to each c_S such that $u \in S$ to be 1 and from p_u to each c_S such that $u \notin S$ to be 3. Other distances are set so as to satisfy the triangle inequality. It is then easy to see that the instance generated is a metric. Now, standard inapproximability result for set cover or for variants such as set coverage imply that it is hard to distinguish between an instance where there is a set of size k covering the universe, and an instance where no set of size k covers more than a $(1-1/e)$ of the universe. Thus, this implies that it is hard to distinguish between an instance of the k-median problem where all points are at distance exactly 1 from their center, and so of cost $|\mathcal{U}|$ and an instance where for any set

of k centers, the number of points at distance 1 is at most $(1-1/e)|\mathcal{U}|$ (and the remaining ones are at distance 3). Hence, hard to distinguish between an instance of cost $|\mathcal{U}|$ and an instance of cost $(1+2/e)|\mathcal{U}|$.

While this reduction yields high inapproximability results in general metric spaces, and so in ℓ_∞-metrics of dimension $\Omega(n)$ through the Fréchet embedding of general metric spaces to ℓ_∞, it seems unrealistic that it can be adapted to ℓ_1- or ℓ_2-metrics, or even $O(\log n)$-dimensional ℓ_∞-metrics.

This is mainly due to the high degree of the hard set cover instances. Indeed, let $d \in \mathbb{N}$ and $\delta \in (0,1)$ be such that $d > \frac{1}{1-\delta^2}$. Then it seems unlikely that we can embed (in any dimension) every d-regular graph into ℓ_2 metric space such that every pair of vertices which had an edge are at distance δ and every non-adjacent pair are at distance 1. The intuition for the previous statement stems from Theorem 5 in [Mae91], which is a special case of the above claim. In other words, the above claim basically says that the maximum gap one can hope for is $poly(1/d)$. This is a constant only when d is a constant.

Another argument for this to be an important roadblock is given recent advances on the parameterized complexity of the problem (parameterized by k), we observe that high dimensional Euclidean space admits a PTAS for k-median and k-means, while arbitrary metric spaces don't (assuming Gap-ETH). The proof that general metric spaces don't admit a fixed-parameter approximation schemes is very similar and so, having an embedding to high dimensional Euclidean space of the above type of instances would contradict Gap-ETH.

1.2.2 Inapproximability in High Dimensions

To obtain our hardness results in high dimensions we will start from the α-vertex coverage problem: given a graph $G(V,E)$ and a parameter k as input, the goal is to distinguish between the following two cases. The Completeness case: There exists $S := \{v_1,\ldots,v_k\} \subseteq V$ such that each edge of E is adjacent to at least one vertex of S, and the Soundness case: For every $S := \{v_1,\ldots,v_k\} \subseteq V$ at most an α fraction of the edges are adjacent to a vertex of S.

This problem will serve for both the discrete and continuous cases (namely the problem where centers have to be picked at specific location and the problem where centers can be picked arbitrarily). We have that the $(0.9292-\varepsilon)$-vertex coverage problem is NP-hard (under unique games conjecture).

Our way of circumventing the problems with embedding the set cover instance (discussed in previous subsubsection) is to reduce from the maximization variant instead of the covering variant as done in previous works [Tre00, GI03, ACKS15, LSW17]. The observation is the that a clustering problems are not covering problems therefore all previous works implicitly paid a factor equal to the degree of the graph in the approximation factor while moving from the vertex cover problem to the clustering problem. By directly using results on the vertex coverage problem, we avoid this. It allows us to look at embeddings where the degree does not inhibit, on the contrary we use the fact that the degree is large to our benefit in some of of our results.

Also, there was no techniques developed in previous works to address all ℓ_p-metrics for the clustering problems. We make an interesting connection to contact dimension of a graph, motivated by recent advances in hardness of approximation in fine-grained complexity [DKL19, KM19]. Elaborating, from the vertex coverage instance $G = (V, E)$ we create the bipartite graph on partite sets V and E where we have an edge $(i, \{j, j'\}) \in V \times E$ if and only if $i = j$ or $i = j'$. Then, we show that embedding this graph so that adjacent vertices are at distance at most β and non-adjacent vertices are at distance at least $\lambda\beta$. From there our inapproximability result follows.

1.2.3 Dimension Reduction

Extending our result to $O(\log n)$-dimensional space while preserving the gap is a challenge for ℓ_1- and ℓ_∞-metric since dimension reduction is very limited for these metrics.

Before we dive into this, let us make the following observation. For the ℓ_2-metric and the k-means objective, the cost of the k-means objective can be expressed as the sum over all clusters of the sum of pairwise distances of points in the cluster divided by the size of the cluster (see Fact 5.6). Thus, it has long been known that dimension reduction using the Johnson-Lindenstrauss lemma preserve the cost of the solutions by a $(1 + \varepsilon)^2$ factor. Hence, we will simply use this to obtain a hardness of approximation for the Euclidean k-means problem in $O(\log n)$ dimension.

Inspired by the recent connections between communication complexity and the hardness of approximation for (geometric) fine-grained and parameterized problems [ARW17, KLM18, Che18] we develop a $O(\log n)$ dimensional embedding technique for all ℓ_p-metrics. An appealing feature of this embedding is that it arises naturally out of the transcript of a (one-way) communication protocol for two-players, where one player is given a vertex of the graph and the other player is given an edge in the graph and the goal is to determine if the vertex covers the edge. We develop non-trivial randomized protocols using algebraic-geometric codes for the aforementioned communication problem, and show how to interpret the transcript to obtain an embedding for both inputs (i.e., the vertex and the edge).

1.3 Organization of the Paper

Section 2 introduces some notations and relevant coding theory concepts and results that will be used throughout the paper. Section 3 discusses graph embedding in ℓ_p-metrics, which form a critical gadget for our hardness results. Section 4 shows our result for the "discrete case", namely when centers have to be picked from a prescribed set. Our results of this section apply to high dimensional spaces, namely when the dimension of the input points is $\Theta(n)$. Section 5 presents our proofs for the "continuous" versions, where centers can be placed at arbitrary locations, also in the case of high dimensional inputs. Sections 6 and 7 presents our dimensionality reduction framework. Finally, Section 8 presents some interesting open problems.

2 Preliminaries

Notations. For any two points $a, b \in \mathbb{R}^d$, the distance between them in the ℓ_p-metric is denoted by $\|a-b\|_p = \left(\sum_{i=1}^{d} |a_i - b_i|^p\right)^{1/p}$. Their distance in the ℓ_∞-metric is denoted by $\|a-b\|_\infty = \max_{i \in [d]} \{|a_i - b_i|\}$, and in the ℓ_0-metric is denoted by $\|a-b\|_0 = |\{i \in [d] : a_i \neq b_i\}|$, i.e., the number of coordinates on which a and b differ. For every $n \in \mathbb{N}$, we denote by $[n]$ the set of first n natural numbers, i.e., $\{1, \ldots, n\}$. We denote by $\binom{[n]}{r}$, the set of all subsets of $[n]$ of size r. Let e_i denote the vector which is 1 on coordinate i and 0 everywhere else. We denote by $\left(\frac{\vec{1}}{2}\right)$, the vector that is $1/2$ on all coordinates.

2.1 Error Correcting Codes

We recall here a few coding theoretic notations. An error correcting code of block length ℓ over alphabet set Σ is simply a collection of codewords $\mathcal{C} \subseteq \Sigma^\ell$. The relative distance between any two points is the fraction of coordinates on which they are different. The relative distance of the code $\mathcal{C}$ is defined to be the smallest relative distance between any pair of distinct codewords in $\mathcal{C}$. The message length of $\mathcal{C}$ is defined to be $\log_{|\Sigma|} |\mathcal{C}|$. The rate of $\mathcal{C}$ is defined as the ratio of its message length and block length.

Theorem 2.1 ([GS96, SAK+01]). *For every prime square q greater than 49, there is a code family over alphabet of size q of positive constant (depending on q) rate and relative distance at least $1 - \frac{3}{\sqrt{q}}$. Moreover, the encoding time of any code in the family is polynomial in the message length.*

The following is an informal argument justifying the existence of the above code family. Fix q a prime square greater than 49.

The authors in [GS96] provide us with a family of curves $\mathsf{C} = \{C_i\}_{i \in \mathbb{N}}$ over $\mathbb{F}_q$ such that for every $\ell \in \mathbb{N}$, we have that C_ℓ has at least ℓ rational points and genus at most $g := 2\ell/\sqrt{q}$. Fix $\ell \in \mathbb{N}$. Let P be a rational point on C_ℓ. Consider the Riemann-Roch space $\mathcal{L}(m \cdot P)$ where $m = \frac{3\ell}{\sqrt{q}}$. This has dimension at least $m + 1 - g = \ell/\sqrt{q} + 1$. Also, any two elements have at most m common zeroes among the rational points of C_ℓ. Pick any set S of ℓ $\mathbb{F}_q$-rational points of C_ℓ that does not contain P. Then the code is given by the evaluations of elements of $\mathcal{L}(m \cdot P)$ at the points of S. The dimension of the code is greater than $\ell/\sqrt{q}$. Therefore the rate is greater than $1/\sqrt{q}$. Also the relative distance of the code is at least $1 - m/\ell = 1 - 3/\sqrt{q}$ as any two codewords agree on at most m coordinates. Finally, the efficient encoding of such a code was given in [SAK+01].

In fact, random codes obtaining weaker parameters than the parameters stated above (see Gilbert-Varshamov bound [Gil52, Var57]) suffice for us[2], but there is no known explicit efficient construction of such codes to the best of our knowledge. It may be possible to use concatenated codes (arising from Reed-Solomon codes) which approach the

[2]To be precise, we need for some infinite increasing sequence $(q_i)_{i \in \mathbb{N}}$ and some increasing function $f : \mathbb{N} \to \mathbb{N}$, a code family over alphabet of size q_i of positive constant (depending on q) rate, relative distance at least $1 - \frac{1}{f(q_i)}$, and efficient encoding.

Gilbert-Varshamov bound in the proofs in this paper instead of the aforementioned algebraic geometric codes.

3 Gadget Constructions via Graph Embeddings

In this section, we first introduce the notion of graph embedding that is of interest to this paper. And then we prove some bounds on the embedding for important ℓ_p-metrics.

Let K_t^r denote the complete r-uniform hypergraph on t vertices (i.e., has all $\binom{t}{r}$ possible hyperedges). Let $\mathcal{I}$ be an operator on hypergraphs which maps every hypergraph to its incidence graph. More formally, for any hypergraph $H(V, E)$ we define $\mathcal{I}(H)$ to be the bipartite graph on partite sets V and E where we have an edge $(i, J) \in V \times E$ in $\mathcal{I}(G)$ (i.e., $J \subseteq V$) if and only if $i \in J$. For every $t, r \in \mathbb{N}$, consider the incidence bipartite graph of the complete hypergraph on t vertices of uniformity (arity) r, which we denote by $H^*(t, r) := \mathcal{I}(K_t^r)$. The vertex set of $H^*(t, r)$ is the partite sets $A^*(t) := [t]$ and $B^*(t, r) := \binom{[t]}{r}$ and (i, J) is an edge in $H^*(t, r)$ if and only if $i \in J$.

We would like to analyze the embedding of $H^*(t, r)$ into ℓ_p-metric spaces for all $p \in \mathbb{R}_{\geq 1} \cup \{\infty\}$.

Definition 3.1 (Gap Realization of a Bipartite graph). *Let $p \in \mathbb{R}_{\geq 1} \cup \{\infty\}$. For any bipartite graph $G = (A \dot\cup B, E)$ and $\lambda \geq 1$, a mapping $\tau : V \to \mathbb{R}^d$ is said to* λ-gap-realize G (in the ℓ_p-metric) *if for some $\beta > 0$, the following holds:*

(i) For all $(u, v) \in E$, $\|\tau(u) - \tau(v)\|_p = \beta$.

(ii) For all $(u, v) \in (A \times B) \setminus E$, we have $\|\tau(u) - \tau(v)\|_p \geq \lambda \cdot \beta$.

Moreover, we require that τ λ-gap-realize G in the ℓ_p-metric efficiently, i.e., there is a polynomial time algorithm (in the size of G) which can compute τ.

We remark here that the above definition is a variant of the notion gap contact dimension introduced in [KM19] in the sense that the authors in [KM19] required that for all distinct u, v both from A or both from B, $\|\tau(u) - \tau(v)\|_p \geq \lambda \cdot \beta$ and for all $(u, v) \in (A \times B) \setminus E$, we have $\|\tau(u) - \tau(v)\|_p > \beta$. They were also interested in the size of the dimension on to which the graph was embedded. Finally, we note that the notion of contact dimension (i.e., with any gap greater than 1) has been studied in literature since the early eighties [Pac80, Mae85, FM86, FM88, Mae91, DKL19].

Definition 3.2 (Gap number). *Let $p \in \mathbb{R}_{\geq 1} \cup \{\infty\}$. For any bipartite graph $G = (A \dot\cup B, E)$, its gap number in the ℓ_p-metric $g_p(G)$ is the largest λ for which there exists a mapping τ that λ-gap-realizes G in a d-dimensional ℓ_p-metric space[3] where $d \leq |A| + |B|$.*

In this paper, we are interested in analyzing $g_p(H^*(t, r))$ for all $t, r \in \mathbb{N}$ and $p \in \mathbb{R}_{\geq 1} \cup \{\infty\}$. We prove the following upper bound[4]:

[3]For all the main results of this paper to hold, we do not require the specified upper bound on the dimension of the mapping realizing the gap number; any finite dimensional realization suffices.

[4]More generally this upper bound holds for any metric (and not necessarily just the ℓ_p-metrics).

Proposition 3.3. *Let $t \geq 3$, $r \geq 2$, and $p \in \mathbb{R}_{\geq 1} \cup \{\infty\}$. If $r < t$ then $g_p(H^*(t,r)) \leq 3$.*

Proof. Let $S = [r-1]$. Let $T = S \cup \{r\}$ and $T' = S \cup \{r+1\}$. Let τ be a λ-gap-realization of $H^*(t,r)$ in the ℓ_p-metric. Then, we have for some $\beta > 0$, that

$$\|\tau(r) - \tau(T)\|_p = \|\tau(r-1) - \tau(T)\|_p = \|\tau(r-1) - \tau(T')\|_p = \beta.$$

But we also have have that $\|\tau(r) - \tau(T')\|_p \geq \lambda\beta$. From triangle inequality this implies $\lambda \leq 3$. □

We can meet the above bound in the ℓ_∞-metric as shown below.

Lemma 3.4. *For all $t \geq 3$ and $r \geq 2$, we have $g_\infty(H^*(t,r)) = 3$.*

Proof. For the ℓ_∞-metric consider the mapping $\tau : A^*(t) \cup B^*(t,r) \to \mathbb{R}^t$ defined as follows. For every $u \in A^*(t)$, we define

$$\tau(u) = e_u + \left(\frac{\vec{1}}{2}\right),$$

and for every $J \in B^*(t,r)$ (i.e., $J \in \binom{[n]}{r}$), we define

$$\tau(J) = \sum_{i \in J} e_i.$$

Fix some $u \in A^*(t)$ and $J \in B^*(t,r)$ such that $u \in J$. Then we have that

$$\eta := \tau(J) - \tau(u) = \left(\sum_{i \in J \setminus \{u\}} e_i\right) - \left(\frac{\vec{1}}{2}\right).$$

Since $\eta \in \{-1/2, 1/2\}^t$, we have that $\|\eta\|_\infty = \|\tau(J) - \tau(u)\|_\infty = 1/2$.

On the other hand if we fix some $u \in A^*(t)$ and $J \in B^*(t,r)$ such that $u \notin J$ then we have that

$$\|\tau(J) - \tau(u)\|_\infty \geq |(\tau(J))_u - (\tau(u))_u| = \frac{3}{2}.$$

Thus we have that τ, 3-gap-realizes $H^*(t,r)$ in the ℓ_∞-metric. Finally, the equality on the gap number follows from Proposition 3.3. □

Next, we consider the ℓ_1-metric and show that we can meet Proposition 3.3 for $r = 2$.

Lemma 3.5. *For all $t \geq 3$ and $r \geq 2$, we have $g_1(H^*(t,r)) \geq \frac{r+1}{r-1}$.*

Proof. For the ℓ_1-metric consider the mapping $\tau : A^*(t) \cup B^*(t,r) \to \{0,1\}^t$ defined as follows. For every $u \in A^*(t)$, we define

$$\tau(u) = e_u,$$

and for every $J \in B^*(t,r)$ (i.e., $J \in \binom{[n]}{r}$), we define

$$\tau(J) = \sum_{i \in J} e_i.$$

Fix some $u \in A^*(t)$ and $J \in B^*(t,r)$ such that $u \in J$. Then we have that

$$\tau(J) - \tau(u) = \sum_{i \in J \setminus \{u\}} e_i \Rightarrow \|\tau(J) - \tau(u)\|_1 = |J| - 1 = r - 1.$$

On the other hand if we fix some $u \in A^*(t)$ and $J \in B^*(t,r)$ such that $u \notin J$ then we have that

$$\|\tau(J) - \tau(u)\|_1 = \left\| \left(\sum_{i \in J} e_i \right) - e_u \right\|_1 = |J| + 1 = r + 1.$$

Thus we have that τ, $\left(\frac{r+1}{r-1}\right)$-gap-realizes $H^*(t,r)$ in the ℓ_1-metric. □

Now we focus our attention to bounding the gap number in the Euclidean metric. We focus on bounding the gap number of $H^*(t,r)$ where $r = 2$, as we only use it later for this fixing of r.

Lemma 3.6. *For all $t \geq 3$, we have $g_2(H^*(t,2)) \geq \frac{2}{\sqrt{(\sqrt{2}-1)^2+1}} \approx 1.848$.*

Proof. Consider the mapping $\tau : A^*(t) \cup B^*(t) \to \{0,1\}^t$ defined as follows. For every $u \in A^*(t)$, we define

$$\tau(u) = \sqrt{2} \cdot e_u,$$

and for every $\{u,v\} \in B^*(t)$, we define

$$\tau(\{u,v\}) = e_u + e_v.$$

Let $i, j, j' \in [t]$ be three distinct numbers. We have

$$\|\tau(i) - \tau(\{i,j\})\|_2 = \|e_i(\sqrt{2}-1) + e_j\|_2 = \sqrt{(\sqrt{2}-1)^2 + 1}, \text{ and}$$

$$\|\tau(i) - \tau(\{j',j\})\|_2 = \|\sqrt{2} \cdot e_i + e_j + e_{j'}\|_2 = 2.$$

This implies τ, $\left(\frac{2}{\sqrt{(\sqrt{2}-1)^2+1}}\right)$-gap realizes $H^*(t,2)$ in the ℓ_2-metric. □

We wrap up our computation of gap numbers by showing that as p grows the gap number of $H^*(t,r)$ in the ℓ_p-metric approaches 3. The proof of the below lemma is very similar to the proof of Lemma 3.4 but we provide it nonetheless for the sake of completeness.

Lemma 3.7. *For all $t \geq 3$ and $r \geq 2$, we have that for every $\varepsilon > 0$ there exists $p \in \mathbb{N}$ such that $g_p(H^*(t,r)) > 3 - \varepsilon$.*

Proof. Fix $t \geq 3$, $r \geq 2$, and $\varepsilon > 0$. Let $p \in \mathbb{N}$ such that $t^{1/p} < 1 + \varepsilon/3$. Consider the mapping $\tau : A^*(t) \cup B^*(t,r) \to \mathbb{R}^t$ defined as follows. For every $u \in A^*(t)$, we define

$$\tau(u) = e_u + \left(\frac{\vec{1}}{2}\right).$$

and for every $J \in B^*(t,r)$ (i.e., $J \in \binom{[n]}{r}$), we define

$$\tau(J) = \sum_{i \in J} e_i.$$

Fix some $u \in A^*(t)$ and $J \in B^*(t,r)$ such that $u \in J$. Then we have that

$$\eta := \tau(J) - \tau(u) = \left(\sum_{i \in J \setminus \{u\}} e_i\right) - \left(\frac{\vec{1}}{2}\right).$$

Since $\eta \in \{-1/2, 1/2\}^t$, we have that $\|\eta\|_p = \|\tau(J) - \tau(u)\|_p = t^{1/p}/2$.

On the other hand if we fix some $u \in A^*(t)$ and $J \in B^*(t,r)$ such that $u \notin J$ then we have that

$$\|\tau(J) - \tau(u)\|_p \geq |(\tau(J))_u - (\tau(u))_u| = \frac{3}{2}.$$

Thus we have that τ, $\left(\frac{3}{t^{1/p}}\right)$-gap-realizes $H^*(t,r)$ in the ℓ_p-metric. Finally note that $\frac{3}{t^{1/p}} > \frac{9}{3+\varepsilon} = 3 - \frac{3\varepsilon}{3+\varepsilon} > 3 - \varepsilon$. □

For most of the results in this paper, we will only use the gap numbers of $H^*(t,r)$ when $r = 2$, and therefore for compactness of statements in the future, we introduce the following.

Definition 3.8. *For all $p \in \mathbb{R}_{\geq 1} \cup \{\infty\}$, we define $\gamma_p = \min_{t \geq 3} g_p(H^*(t,2))$.*

Finally, we conclude this section by showing a 'hereditary' property of our embedding which will be invoked for all applications.

Proposition 3.9. *Let G be a r-uniform hypergraph on $t \geq 3$ vertices and let $H := \mathcal{I}(G)$. Let τ be a λ-gap realization of $H^*(t,r)$ in the ℓ_p-metric. Then τ restricted to the vertices of H is a λ-gap realization of H in the ℓ_p-metric. In particular, for $r = 2$, there exists a mapping τ^* which is a γ_p-gap realization of H in the ℓ_p-metric.*

We skip the proof of the above proposition as it follows in a straightforward manner from Definition 3.1.

4 Inapproximability of k-means and k-median with Candidate Centers in High Dimensions

In this section, we prove Theorem 1.2 but in high dimensions by a reduction from the gap vertex coverage problem. First, we define the gap vertex coverage problem.

Let $G(V, E)$ be a graph. Let $S \subseteq V$. We define the cover of S, denoted by $\mathsf{cov}(S)$ as follows:

$$\mathsf{cov}(S) = \{e \in E \mid \exists v \in S \text{ such that } v \in e\}.$$

Definition 4.1 (α-vertex coverage)**.** *In the α-vertex coverage problem, we are given a graph $G(V, E)$ and a parameter k as input. We would like to distinguish between the following two cases:*

- **Completeness***: There exists $S := \{v_1, \ldots, v_k\} \subseteq V$ such that $\mathsf{cov}(S) = E$.*
- **Soundness***: For every $S := \{v_1, \ldots, v_k\} \subseteq V$ we have $|\mathsf{cov}(S)| \leq \alpha \cdot |E|$.*

We recall that the minimum degree of a graph is said to be $d_{\min} \in \mathbb{N}$ if every vertex in the graph has degree at least $d_{\min}$. Strong inapproximability results were given for the vertex coverage problem in[5] [AKS11]. Recently, Austrin and Stanković provided the tight inapproximability result, which is stated below.

Theorem 4.2 (Austrin and Stanković [AS19])**.** *There is some $\varepsilon > 0$ and $d_0 \in \mathbb{N}$, such that for all $d_{\min} > d_0$, assuming the unique games conjecture, deciding an instance (G, k) of $(0.9292 - \varepsilon)$-vertex coverage problem on minimum degree $d_{\min}$ graphs is* NP*-hard.*

We remark here that [AS19] computed the inapproximability factor up to 3 decimal places, and the above hardness of approximation factor follows from additional computation. Also note that in [AS19] the hardness is not shown for minimum degree $d_{\min}$ graphs[6] but if we look at the removal of vertex weights step in Section 4 of [AKS11] then we can take large enough number of copies ($> d_{\min}$) of each vertex (proportional to its weight) and this will ensure the theorem as stated above.

Another important remark is that the hardness results of [AKS11, AS19] are for multigraph instances of the vertex coverage problem. However, in this paper, we treat that the hard instances of Theorem 4.2 are simple graphs for the sake of brevity. This assumption is reasonable, because in all our reductions, we realize every edge of the graph as a point in space, and in the case of a multigraph, we realize two edges with the same pair of end points, as two distinct points, where one of the points is just a slightly perturbed version of the other point. It is then clear that all points which correspond to edges with the same pair of end points, will be in the same cluster, in the optimal solution.

Next we define for every $p \in \mathbb{R}_{\geq 1} \cup \{\infty\}$,the quantities $\zeta_1(p)$ and $\zeta_2(p)$ as follows:

$$\zeta_1(p) := 0.9292 + (\gamma_p \cdot 0.0708) \text{ and } \zeta_2(p) := 0.9292 + (\gamma_p^2 \cdot 0.0708).$$

Again notice that $\zeta_1(1) = 1.1416$, $\zeta_2(1) = 1.5664$, $\zeta_1(2) \approx 1.06$, $\zeta_2(2) \approx 1.1709$, and as $p \to \infty$, we have $\zeta_1(p) \to \zeta_1(\infty) = \zeta_1(1)$ and $\zeta_2(p) \to \zeta_2(\infty) = \zeta_2(1)$.

[5]The result is implicit in [AKS11], and is explicitly written in [Man19].

[6]We require the hard instances of gap vertex coverage problem to have this additional minimum degree requirement only for proving our inapproximability results of clustering objectives in the continuous case (i.e., Theorems 1.3 and1.4), and do not need it to prove our hardness of approximation results in the discrete case (i.e., Theorem 1.2).

Now, we state our inapproximability results for k-means and k-median in high dimensions.

Theorem 4.3 (k-means with candidate centers in $n^{O(1)}$ dimensional ℓ_p-metric space)**.** *Let $p \in \mathbb{R}_{\geq 1} \cup \{\infty\}$. Assuming the unique games conjecture, given a point-set $\mathcal{P} \subset \mathbb{R}^m$ of size n (and $m = poly(n)$), a collection $\mathcal{C}$ of m candidate centers in $\mathbb{R}^m$, and a parameter k as input, it is* NP*-hard to distinguish between the following two cases:*

- **Completeness***: There exists $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and $\sigma : \mathcal{P} \to \mathcal{C}'$ such that*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_p^2 \leq \beta^2 n,$$

- **Soundness***: For every $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and every $\sigma : \mathcal{P} \to \mathcal{C}'$ we have:*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_p^2 \geq \zeta_2(p) \cdot \beta^2 n,$$

for some constant $\beta > 0$.

Theorem 4.4 (k-median with candidate centers in $n^{O(1)}$ dimensional ℓ_p-metric space)**.** *Let $p \in \mathbb{R}_{\geq 1} \cup \{\infty\}$. Assuming the unique games conjecture, given a point-set $\mathcal{P} \subset \mathbb{R}^m$ of size n (and $m = poly(n)$), a collection $\mathcal{C}$ of m candidate centers in $\mathbb{R}^m$, and a parameter k as input, it is* NP*-hard to distinguish between the following two cases:*

- **Completeness***: There exists $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and $\sigma : \mathcal{P} \to \mathcal{C}'$ such that*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_p \leq \beta n,$$

- **Soundness***: For every $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and every $\sigma : \mathcal{P} \to \mathcal{C}'$ we have:*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_p \geq \zeta_1(p) \cdot \beta n,$$

for some constant $\beta > 0$.

Proof of Theorems 4.3 and 4.4 . Starting from a hard instance of $(0.9292 - \varepsilon)$-vertex coverage problem $G = (V, E)$ which is guaranteed by Theorem 4.2, we create an instance of the k-means, or of the k-median problem using the embedding given in Proposition 3.9 as follows. Let τ be the embedding of $\widetilde{G} := \mathcal{I}(G)$ prescribed by Proposition 3.9. We think of $\widetilde{G}$ as $\widetilde{G}(V \cup B, \mathcal{E})$ where V is simply the vertex set of the $(0.9292 - \varepsilon)$-vertex coverage instance, and B is obtained by defining a vertex $b_{i,j}$ for each edge (u_i, u_j) of the $(0.9292 - \varepsilon)$-vertex coverage instance, and $\mathcal{E}$ is obtained by defining an edge from each vertex $u_i \in V$ to each vertex $b_{i,j} \in B$.

The k-median or k-means instance consists of the set of candidate centers $\mathcal{C}$ and the set of points to be clustered $\mathcal{P}$ defined as follows:

$$\mathcal{P} = \{\tau(v) \mid v \in B\} \text{ and } \mathcal{C} = \{\tau(u) \mid u \in V\}.$$

We now analyze the k-means and k-median cost of the instance. Consider the completeness case first.

Completeness. In this case, we know that there is a set $S \subset V$ such that $|S| = k$ and S covers all the edges of the $(0.9292 - \varepsilon)$-vertex coverage instance G. We focus on the set of centers $\mathcal{C}'$ induced by S, namely

$$\mathcal{C}' = \{\tau(u_i) \mid u_i \in S\} \subseteq \mathcal{C}.$$

Since each edge (u_i, u_j) is adjacent to at least one element of S, we have that for every $\tau(b_{i,j})$, the following holds:

$$\min_{c \in \mathcal{C}'} \|\tau(b_{i,j}) - c\|_p^2 = \beta^2 \text{ and } \min_{c \in \mathcal{C}'} \|\tau(b_{i,j}) - c\|_p = \beta,$$

for some $\beta > 0$. The k-means cost of the overall instance is thus $\beta^2 \cdot |\mathcal{P}|$, while the k-median cost is $\beta \cdot |\mathcal{P}|$. Finally, we turn to the soundness analysis.

Soundness. Consider any set of centers $\mathcal{C}' = \{c_1, \ldots, c_k\} \subset \mathcal{C}$ that is optimal for the k-median or k-means objective. Let $S := \{v_1, \ldots, v_k\}$ be the set of vertices corresponding to the centers of $\mathcal{C}'$, namely

$$S = \{v \in V \mid \tau(v) \in \mathcal{C}'\}.$$

By the assumptions of the soundness case, S covers at most $(0.9292 - \varepsilon)|E|$ number of edges of G. For each such edge, $e = (u_i, u_j)$, we have that the contribution of $\tau(e)$ to the k-means cost is exactly β^2, and to the k-median cost is exactly β. By the definition of the gadget τ, we have that for any other edge $e = (u_i, u_j)$ that is not covered by S, the contribution of $\tau(e)$ to the k-median and k-means cost is respectively $\gamma_p \cdot \beta$ and $\gamma_p^2 \cdot \beta^2$. Therefore, the optimal solution w.r.t. k-median objective has cost at least $\zeta_1(p) \cdot \beta \cdot |\mathcal{P}|$, and optimal solution w.r.t. k-means objective has cost at least $\zeta_2(p) \cdot \beta^2 \cdot |\mathcal{P}|$, as claimed. □

We would like to conclude this section by remarking that the hardness of approximation results for the ℓ_∞-metric given above are strictly weaker than the known hardness of approximation factors for this metric [GK99]. By a straightforward application of the Fréchet embedding[7] to the constructions in [GK99], we obtain the NP-hardness of approximating k-means (resp. k-median) to a factor better than $1 + 8/e$ (resp. $1 + 2/e$). These inapproximability factors are much higher than the ones given in Theorems 4.3 and 4.4. However, our main contribution as far as the ℓ_∞-metric is concerned is to obtain the same inapproximability factor as [GK99] but in low dimensions (i.e., $O(\log n)$ dimensions). This result is proven in Section 7.3.

[7]The Fréchet embedding maps n points in any metric into the ℓ_∞-metric with polynomial in n blowup in the dimension such that all pairwise distances are preserved.

5 Inapproximability of k-median and k-means without Candidate Centers in High Dimensions

In this section, we prove Theorems 1.3 and 1.4 but in high dimensions. In particular, in Section 5.1 we show our inapproximability results for the k-median and k-means objectives without candidate centers in the Hamming metric. In Section 5.2, we show our inapproximability result for the k-median objective without candidate centers in the ℓ_1-metric. Finally, in Section 5.3, we show our inapproximability result for the k-means objective without candidate centers in the ℓ_2-metric.

5.1 Inapproximability in Hamming metric

In this subsection, we prove our inapproximability results for the k-median and k-means objectives without candidate centers in the Hamming metric. Our proof considers the same reduction from the gap vertex coverage problem as described in the proofs of Theorem 4.3 and 4.4, but performs the completeness and soundness analysis for the case where there is no candidate centers set given as part of the input.

Theorem 5.1 (k-means without candidate centers in $n^{O(1)}$ dimensional Hamming metric space)**.** *Assuming the unique games conjecture, given a point-set $\mathcal{P} \subset \{0,1\}^m$ of size n (and $m = poly(n)$) and a parameter k as input, it is* NP*-hard to distinguish between the following two cases:*

- **Completeness***: There exists $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \{0,1\}^m$ and $\sigma : \mathcal{P} \to \mathcal{C}'$ such that*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_0^2 \leq n,$$

- **Soundness***: For every $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \{0,1\}^m$ and every $\sigma : \mathcal{P} \to \mathcal{C}'$ we have:*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_0^2 \geq 1.21 \cdot n.$$

Theorem 5.2 (k-median without candidate centers in $n^{O(1)}$ dimensional Hamming metric space)**.** *Assuming the unique games conjecture, given a point-set $\mathcal{P} \subset \{0,1\}^m$ of size n (and $m = poly(n)$) and a parameter k as input, it is* NP*-hard to distinguish between the following two cases:*

- **Completeness***: There exists $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \{0,1\}^m$ and $\sigma : \mathcal{P} \to \mathcal{C}'$ such that*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_0 \leq n,$$

- **Soundness***: For every $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \{0,1\}^m$ and every $\sigma : \mathcal{P} \to \mathcal{C}'$ we have:*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_0 \geq 1.07 \cdot n.$$

Proof of Theorems 5.1 and 5.2. Starting from an instance of $(0.9292 - \varepsilon)$-vertex coverage problem $G = (V, E)$, we create an instance of the k-means, or of the k-median problem using the 3-gap realization mapping τ for the ℓ_1-metric given in Lemma 3.5 (by setting $r = 2$) as follows. First note that τ maps points to the Boolean hypercube and for a pair of points on the Boolean hypercube the distances in the Hamming and ℓ_1-metric are the same. Let τ be the embedding of $\widetilde{G} := \mathcal{I}(G)$ prescribed by Lemma 3.5. We think of $\widetilde{G}$ as $\widetilde{G}(V \cup B, \mathcal{E})$ where V is simply the same set of vertices as the vertex set of vertices of the vertex coverage instance, B is obtained by defining a vertex $b_{i,j}$ for each edge (u_i, u_j) of the vertex coverage instance, and $\mathcal{E}$ is obtained by defining an edge from each vertex $u_i \in V$ to each vertex $b_{i,j} \in B$.

The k-median or k-means instance without candidate centers is just the set of points $\mathcal{P} := \{\tau(v) \mid v \in B\}$ that we would like to cluster. In particular notice that for all $v \in B$ we have $\|\tau(v)\|_0 = 2$. We now analyse the k-means and k-median cost of the instance. Consider the completeness case first.

Completeness. In that scenario, pick a vertex coverage V^* of the instance and focus on the set of centers C^* induced by V^*, namely $C^* = \{\tau(u_i) \mid u_i \in V^*\}$. Since by definition of vertex coverage, each edge (u_i, u_j) is adjacent to at least one element of V^* and so for each point $\tau(b_{i,j})$ we have that $\min_{c \in C^*} \|\tau(b_{i,j}) - c\|_0^2 = 1 = \min_{c \in C^*} \|\tau(b_{i,j}) - c\|_0$. The k-means and k-median cost of the overall instance is at most $|E|$. Thus, let's turn to the soundness case.

Soundness. Consider any set of centers $C^* = \{c_1, \ldots, c_k\} \subseteq \{0,1\}^{|V|}$ that is optimal for the k-median or k-means objective. Fix some arbitrary $i \in [k]$. Note that if $\|c_i\|_0 \geq 4$ then, $\|c_i - \tau(u)\| \geq 2$ for any $u \in B$, and thus we could replace c_i by the all zeroes vector and the cost of k-means or k-median would not increase. Therefore we assume all the centers have Hamming weight at most 3. We partition C^* into C_0, C_1, C_2, and C_3 where $c \in C^*$ belongs to C_j if the Hamming weight of c is j. Consider an optimal classification $\sigma : \mathcal{P} \to C^*$. For every point $c \in C^*$ let $T_c^\sigma \subseteq B$ be defined as follows:

$$T_c^\sigma = \{u \in B \mid \sigma(\tau(u)) = c\}.$$

We propose the following claim.

Claim 5.3. *Given an optimal classification σ we can construct an optimal classification σ^* (which might be same as σ) such that for any $c \in C_3$ we have $|T_c^{\sigma^*}| \leq 3$ and for any $c \in C_2$ we have $|T_c^{\sigma^*}| \leq 1$.*

Before we prove the above claim, we see how it completes the proof. For every $c \in C_1$ if its 1 is on coordinate i we associate it with the vertex i in G. Now we partition $T_c^{\sigma^*}$ into $Y_c^{\sigma^*}$ and $N_c^{\sigma^*}$ where for any $u \in B$ such that $\sigma^*(\tau(u)) = c$ we have that $u \in Y_c^{\sigma^*}$ if $c \in u$ (think of c as the vertex in G) and $u \in N_c^{\sigma^*}$ otherwise. By definition of the soundness case, we have that $\sum_{c \in C_1} |Y_c^{\sigma^*}|$ is at most $(0.9292 - \varepsilon)|E|$. Notice that there are at most $3|C_3| + |C_2|$ edges which are not assigned to a center in $C_1 \cup C_0$. We upper bound $|C_3|, |C_2$ by $|V|$ and thus we have that there are at most $4|V|$ edges which are not

assigned to a center in $C_1 \cup C_0$. If an edge is assigned a center in C_0 then its distance from the center is 2. If an edge is assigned a center in $c \in C_1$ and it is contained in $Y_c^{\sigma^*}$ then its distance from the center is 1; but if it is contained in $N_c^{\sigma^*}$ then its distance from the center is 3. Therefore we have that there are at least $(0.0708 + \varepsilon)|E| - 4|V|$ edges that are distance at least 2 from their allocated center according to the clustering σ^*. Notice that in Theorem 4.2 we can choose the minimum degree of G to be as large a constant as we want. We choose it to be greater than $8/\varepsilon$. In this case we have that $|E| \geq 4|V|/\varepsilon$. Therefore, the optimal solution w.r.t. k-median objective has cost at least $(2(0.0708) + 0.9292) \cdot |E| = 1.0708 \cdot |E|$, and optimal solution w.r.t. k-means objective has cost at least $(4(0.0708) + 0.9292) \cdot |E| = 1.2124 \cdot |E|$, as claimed. □

We can thus conclude the proof of the above theorem by proving Claim 5.3.

Proof of Claim 5.3. Now we show how to construct σ^* from σ. Consider $c \in C_3$. Let the three coordinates where c is 1 be i, j, j'. We think of i, j, j' as vertices in G. Let $F = \{(i, j), (i, j'), (j, j')\}$. For any edge not in F its distance to C_3 is 3 from c. If any of the points corresponding to i, j, j' under τ was picked in our set of centers then, we could replace c by another point in $\tau(i), \tau(j), \tau(k)$. If this is not the case then only the edges in $E \cap F$ would be at distance 1 from c, and for the rest we could choose some point in C_1 or C_0. Thus we would obtain a new optimal classification in which points assigned to c would be at most 3. Similar (and simpler) argument holds for $c \in C_2$ as any two edges under τ are at distance 2. □

5.2 Inapproximability of k-median in ℓ_1-metric

In this subsection, we prove Theorem 1.4 but in high dimensions. At a high level, our proof simply considers the hard instances built in the Hamming metric in Theorem 5.2, and notes that the optimum cluster centers for those instances in the ℓ_1-metric must have coordinate entries in $\{0, 1\}$.

Theorem 5.4 (k-median without candidate centers in $n^{O(1)}$ dimensional ℓ_1-metric space)**.** *Assuming the unique games conjecture, given a point-set $\mathcal{P} \subset \{0,1\}^m$ of size n (and $m = poly(n)$) and a parameter k as input, it is* NP*-hard to distinguish between the following two cases:*

- **Completeness***: There exists $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathbb{R}^d$ and $\sigma : \mathcal{P} \to \mathcal{C}'$ such that*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_1 \leq n,$$

- **Soundness***: For every $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathbb{R}^d$ and every $\sigma : \mathcal{P} \to \mathcal{C}'$ we have:*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_1 \geq 1.07 \cdot n.$$

Proof. The proof follows from a simple observation and mimicking the proof of Theorem 5.2. Recall first that given a set of z integers $X = \{x_1, \ldots, x_z\}$, we have that a median

of X is a point x^* that minimizes $\sum_{i=1}^{z} |x^* - x_i|$. Note that if X contains an even number of points, then the median is not unique, nonetheless, there is always at least one point of X minimizing $\sum_{i=1}^{z} |x^* - x_i|$. We refer to these points as the discrete medians.

Thus, consider a set of points P in a dimensional ℓ_1-metric space. The points p^* that minimizes $\sum_{p \in P} \|p - p^*\|_1$ is therefore the point p^* whose ith coordinate is the median of the ith coordinates of the points in P.

Hence, consider an instance of the k-median problem in Hamming metric as defined in the proof of Theorem 5.2 and apply the same construction to obtain an instance in ℓ_1. We have that for this instance all the coordinates of the points to be clustered are in $\{0, 1\}$. Thus, for any subset (i.e. cluster) of the points of the instance, an optimal center of the set is such that its ith coordinate is the median of a set of values in $\{0, 1\}$. From the above discussion, we conclude that assuming that the ith coordinate is also in $\{0, 1\}$ is without loss of generality. It follows that for any clustering, we can assume that the centers induced by the partition have coordinates in $\{0, 1\}$.

Therefore, the rest of the proof follows by applying the same reasonning than in the proof of Theorem 5.2 since the instance created behaves in ℓ_1 metric like the instance described in the proof of Theorem 5.2 in Hamming metric. □

5.3 Inapproximability of k-means in Euclidean metric

In this section, we prove Theorem 1.3 but in high dimensions.

Theorem 5.5 (k-means without candidate centers in $n^{O(1)}$ dimensional ℓ_2-metric space)**.** *Assuming the unique games conjecture, given a point-set $\mathcal{P} \subset \{0, 1\}^m$ of size n (and $m = poly(n)$) and a parameter k as input, it is* NP*-hard to distinguish between the following two cases:*

- **Completeness***: There exists $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathbb{R}^d$ and $\sigma : \mathcal{P} \to \mathcal{C}'$ such that*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_2^2 \leq \beta n,$$

- **Soundness***: For every $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathbb{R}^d$ and every $\sigma : \mathcal{P} \to \mathcal{C}'$ we have:*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_2^2 \geq 1.07 \cdot \beta n,$$

for some constant $\beta > 0$.

Proof. Let $\varepsilon > 0$ and $(G = (V, E), k)$ be an instance of the $(0.9292 - \varepsilon)$-vertex coverage problem on graph of minimum degree at least $\varepsilon_0^{-1} := \frac{20}{\varepsilon^4}$. By Theorem 4.2, for some $\varepsilon > 0$, we have that deciding such an instance is NP-hard, assuming the unique games conjecture.

We build a set of points $\mathcal{P}$ as follows[8]: for each edge e_{u_i,u_j}, we create a point $p_{i,j}$

[8]The construction described here is equivalent to using the 1.848-gap realization mapping τ for the ℓ_2-

whose ith and jth coordinates are both 1 and whose remaining coordinates are all 0. We say that $p_{i,j}$ is the point corresponding to edge e_{u_i,u_j}. We thus have the following fact:

Fact 5.6. *Consider two edges e, f. If $e = (u_i, u_j)$ and $f = (u_i, u_\ell)$ then $\|p_{i,j} - p_{i,\ell}\|_2^2 = 2$. If $e = (u_i, u_j)$ and $f = (u_r, u_\ell)$, where $r, \ell \notin \{i, j\}$ then $\|p_{i,j} - p_{r,\ell}\|_2^2 = 4$.*

We now prove the completeness and soundness cases.

Completeness. In the completeness case, we show that the cost of the optimal solution is at most $|E| - |V|/2$. By definition, there exists a vertex cover S of size $k = n/2$ of the instance. Define a partition of the edge set into k parts by *assigning* each edge it to one of its extremity that is in S, let $\{C_1, \ldots, C_k\}$ be the partition induced by the assignment. Note that since there exists such a vertex cover, such a partition is indeed possible. We now bound the k-means cost of solution $\{C_1, \ldots, C_k\}$. We claim that for each cluster C_i, $\text{cost}(C_i) = m_i - 1$, where $m_i = |C_i|$.

Indeed, let v_j be the vertex covering all edges of C_i. Observe first that the jth coordinate of the centroid of C_i is 1. The remaining coordinates of the centroid of C_i are 0 except for m_i of them which are $1/m_i$.

Then, for a given edge (v_j, v_ℓ), the k-means cost of the corresponding point is $(1 - 1/m_i)^2 + (m_i - 1)(1/m_i)^2 = 1 - 2/m_i + m_i^2 + 1/m_i - 1/m_i^2$ which is $1 - 1/m_i$. Summing up over all the m_i edges of C_i yields that $\text{cost}(C_i) = m_i - 1$. Therefore, the total k-means cost of the clustering is $m - n/2$.

Soundness In the soundness case, we show that the optimal k-means cost is at least $(1.114 - 2\varepsilon)|E| - |V|$. We will use the following classic fact about the k-means objective.

Fact 5.7. *Given a clustering $\{C_1, \ldots, C_k\}$, the k-means cost is exactly*

$$\sum_{i=1}^{k} \frac{1}{2|C_i|} \sum_{p \in C_i} \sum_{q \in C_i} \|p - q\|_2^2$$

Now, consider an optimal clustering $\{C_1, \ldots, C_k\}$ of the instance in the soundness case. For each cluster C_i, we define the graph G_i to be the subgraph of the graph G induced by the edges whose corresponding points are in C_i. We let Δ_i be the maximum degree in G_i. We have the following claim.

Claim 5.8. *For any cluster C_i such that $|C_i| \geq 10/\varepsilon^3$, we have cost$(C_i) \geq 2(1 - \varepsilon)|C_i| - (1 + \varepsilon)\Delta_i$.*

Assume Claim 5.8 is true for a moment. Then, the proof of the lemma can be completed as follows. First, observe that since the number of clusters is $k = n/2$ and the graph G has at least $10n/\varepsilon^4$, the total number of edges in clusters C_i such that $|C_i| < 10/\varepsilon^3$ is at most $\varepsilon m/2$. Let's assume the cost for these edges is 0 and let's focus on the cost of

metric given in Lemma 3.6 in the following way. The k-median or k-means instance without candidate centers is just the set of points $\mathcal{P} := \{\tau(e) \mid e \in E\}$ that we would like to classify. In particular notice that for all $e \in E$ we have $\|\tau(e)\|_2^2 = 2$.

clusters of size at least $10/\varepsilon^3$, let k' be the number of such clusters. Summing up over all such clusters we have that the total k-means cost is at least $\sum_{i=1}^{k'} 2(1-\varepsilon)|C_i| - (1+\varepsilon)\Delta_i \geq (2-3\varepsilon)m - (1+\varepsilon)\sum_{i=1}^{k'}\Delta_i$. Then, to provide a lower bound on Δ_i, consider the set S obtained by picking a vertex of degree Δ_i from each G_i. This set has size at most $n/2$ and so by definition of the soundness case the sum of the degrees of the vertices in S in G is at most $(0.9292-\varepsilon)m$. It follows that the cost of the optimal solution is at least $2m - ((1+\varepsilon)0.9292 - 4\varepsilon)m$ which is at least $(1.0708 - O(\varepsilon)) \cdot m$ as stated. □

We can thus conclude the proof of the above theorem by proving Claim 5.8.

Proof of Claim 5.8. Consider a cluster C_i such that $|C_i| \geq 10/\varepsilon^3$. Consider an edge $e = (u_\ell, u_j)$ whose corresponding point p is in C_i. By Facts 5.7 and 5.6 we have that

$$\frac{1}{2|C_i|}\sum_{q\in C_i}\|p-q\|_2^2 = \frac{1}{2|C_i|}\left(2(d_{i,\ell}+d_{i,j}-2)+4(|C_i|-d_{i,\ell}-d_{i,j}+2)\right),$$

where $d_{i,\ell}, d_{i,j}$ are the degrees of vertices u_ℓ, u_j respectively in G_i. Now, summing up over all edges in C_i this gives a total cost for the cluster C_i of

$$\frac{|C_i|}{2|C_i|}(4|C_i|) - \frac{2}{2|C_i|}\sum_{e=(u_\ell,u_j)\in C_i}(d_{i,\ell}+d_{i,j}-2)$$

which is

$$2|C_i| + 2 - \frac{1}{|C_i|}\sum_{u_j} d_{i,j}^2.$$

We now need to provide an upper bound on $\frac{1}{|C_i|}\sum_{u_j} d_{i,j}^2$. First, consider the set S of vertices u_j such that $d_{i,j} < \varepsilon|C_i|$ and let m_i^0 be the number of edges with at least one extremity in S. We have that

$$\frac{1}{|C_i|}\sum_{u_j\in S} d_{i,j}^2 \leq \frac{\varepsilon|C_i|}{|C_i|}\sum_{u_j\in S} d_{i,j} \leq 2\varepsilon m_i^0. \tag{1}$$

We then bound $\sum_{u_j\notin S} d_{i,j}^2$. Let S' be the set of vertices with degree larger than $\varepsilon|C_i|$ in G_i. Moreover, let m_i^1 be the set of edges with both extremities in S'. We start by arguing that $m_i^1 < \varepsilon m$.

We have that $\sum_{u_j\in S'} d_{i,j} \leq m_i^1 + |C_i|$ and so, there exists a vertex u_j in S' such that $d_{i,j} \leq (m_i^1 + |C_i|)/|S'|$. Thus, since $d_{i,j} \in S'$, we have that $d_{i,j} > \varepsilon|C_i|$ and so $(m_i^1 + |C_i|)/|S'| \geq \varepsilon|C_i| \geq \varepsilon m_i^1$. This implies that $|S'|m_i^1 \leq (m_i^1 + |C_i|)\varepsilon^{-1}$. Now, assume towards contradiction that $m_i^1 > \varepsilon|C_i|$. Then $|S'| < \varepsilon^{-1}(\varepsilon^{-1}+1)$. Combining this with the fact that $\varepsilon|C_i| \leq (m_i^1 + |C_i|)/|S'|$, we have that $|C_i|(|S'|-1) \leq m_i^1 \leq \frac{|S'|(|S'|-1)}{2}$. Hence, $|C_i| \leq |S'|/2 \leq \varepsilon^{-1}(\varepsilon^{-1}+1)$ and so $|C_i| \leq 10\varepsilon^{-3}$, a contradiction. Therefore $m_i^1 \leq \varepsilon|C_i|$.

We can now bound $\sum_{u_j\notin S} d_{i,j}^2$. We have

$$\sum_{u_j\notin S} d_{i,j}^2 \leq \Delta_i \sum_{u_j\notin S} d_{i,j} \leq \Delta_i(1+\varepsilon)|C_i|. \tag{2}$$

Finally, combining Equations 1 and 2 we deduce that $\frac{1}{|C_i|}\sum_{u_j} d_{i,j}^2 \leq 2\varepsilon|C_i| + (1+\varepsilon)\Delta_i$. Therefore, $\text{cost}(C_i) \geq 2(1-\varepsilon)|C_i| - (1+\varepsilon)\Delta_i$. □

5.3.1 k-means in Euclidean Metric over Reals in Low Dimensions

In this subsection, we prove Theorem 1.3 albeit over real vectors.

Theorem 5.9 (k-means in Euclidean metric in $O(\log n)$ dimensions without Candidate Centers over Reals)**.** *Let ε be an arbitrarily small constant. Assuming the unique games conjecture, given a point-set $\mathcal{P} \subset \mathbb{R}^d$ of size n (and $d = O(\log n)$) and a parameter k as input, it is* NP*-hard to distinguish between the following two cases:*

- **Completeness***: There exists $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathbb{R}^d$ and $\sigma : \mathcal{P} \to \mathcal{C}'$ such that*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_2^2 \leq \beta n,$$

- **Soundness***: For every $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathbb{R}^d$ and every $\sigma : \mathcal{P} \to \mathcal{C}'$ we have:*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_2^2 \geq (1.07 - \varepsilon) \cdot \beta n,$$

for some constant $\beta > 0$.

Proof. By Theorem 5.5, we have that given a set of points $\mathcal{P} \in \mathbb{R}^{O(n)}$, it is hard to distinguish between the following cases:

- ***Completeness***: There exists $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathbb{R}^d$ and $\sigma : \mathcal{P} \to \mathcal{C}'$ such that

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_2^2 \leq \beta n,$$

- ***Soundness***: For every $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathbb{R}^d$ and every $\sigma : \mathcal{P} \to \mathcal{C}'$ we have:

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_2^2 \geq 1.07 \cdot \beta n,$$

The proof of Theorem 5.9 follows from the Johnson-Lindenstrauss lemma and the following well-known observation. Given a set of n points in $\mathbb{R}^d$, we have that by Fact 5.7 the k-means cost of a given partition $\{C_1, \ldots, C_k\}$ can be expressed as $\sum_{i=1}^{k} \frac{1}{2|C_i|} \sum_{x,y \in C_i} \|x - y\|_2^2$. Thus, applying the Johnson-Lindenstrauss lemma with target dimension $O(\log n/\varepsilon^5)$ for small enough ε, yields an instance where the k-means cost of any clustering C is within a factor $(1+\varepsilon)$ of the k-means cost of C in the original d-dimensional instance. It follows that the gap is preserved up to a $(1+\varepsilon)$ factor and the theorem follows.

Note that this can be made deterministic (for example, see the result of Engebretsen et al. [EIO02]). □

6 Embedding via Communication Protocols

In this section, we introduce our (hyper)graph embedding technique, which will enable us to prove the same hardness results as in Sections 4 and 5 but for point-sets in $O(\log n)$ dimensions.

6.1 One-Way Communication Model and Protocols

In this subsection, we first introduce a communication model known in literature as the one-way communication model.

The two-player One-Way Communication (OWC) model was introduced by Yao [Yao79] and has been extensively studied in literature [KN97].

One-Way Communication Model. Let $\mathcal{X}$ and $\mathcal{Y}$ be two finite sets. Let $f : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$. In the two-player one-way communication model, we have Alice and Bob each with an input $x \in \mathcal{X}$ and $y \in Y$ respectively, and the communication task is for Bob to determine if $f(x,y) = 1$. In this model, only Alice is allowed to send messages to Bob. In the randomized setting, we allow the players to jointly toss some random coins before sending messages, i.e., we allow public randomness. Moreover, we assume that the sets $\mathcal{X}, \mathcal{Y}$ are public knowledge.

Next, we introduce the notion of OWC protocols, which are in a nutshell one-round randomized protocols where the players are in a computationally bounded setting.

OWC **Protocols.** Let π be a communication protocol for a problem in the OWC model. We say that π is a (r, μ, α, s)-OWC protocol if the following holds:

- The protocol is one-round with public randomness, i.e., the following actions happen sequentially:
  1. The players receive their inputs.
  2. The players jointly toss r random coins.
  3. Alice on seeing the randomness (i.e. results of r coin tosses) deterministically sends an μ-bit message to Bob.
  4. Based on the μ bits sent from Alice and randomness r, Bob outputs accept or reject.
- The protocol has completeness 1 and soundness s, i.e.,
  - If $f(x,y) = 1$, then Bob always accepts.
  - If $f(x,y) = 0$, then Bob accepts with probability at most s.
- We have that the expected number of distinct messages that Alice could send (on randomness r) which Bob would accept is α, where the expectation is over the randomness r.
- The players are computationally bounded, i.e., all of them perform all their computations in poly$(|\mathcal{X}| + |\mathcal{Y}|)$-time.

In a (r, μ, α, s)-OWC protocol, we refer to r as the randomness complexity of the protocol, μ as the message complexity of the protocol, α as the acceptance complexity of the protocol, and s as the soundness of the protocol. We note here that while the randomness complexity and message complexity are standard measures of interest in literature, the acceptance complexity is non-standard but the measure is important for our embedding later. We note here that the acceptance complexity is closely related to the free bit complexity measure studied in PCP literature [BGS98].

Definition 6.1 (c-left bounded functions)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be two finite sets and $c \in \mathbb{N}$. Let $f : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$. For every $y \in \mathcal{Y}$, let $S_y = \{a \in \mathcal{X} \mid f(a,y) = 1\}$. Then f is said to be c-left bounded if for every $y \in \mathcal{Y}$, we have $|S_y| = c$.*

For every c-left bounded f, there is a trivial (deterministic) $(0, \log|\mathcal{X}|, c, 0)$-OWC protocol. We would like to use randomness to do better on the message complexity.

Theorem 6.2. *Let $\mathcal{X}$ and $\mathcal{Y}$ be sets of size m and n respectively. Let $f : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ be a c-left bounded function. For every prime square $q \gg c^4$, there is a $(O_q(1) + \log\log m, \lceil \log_2 q \rceil, \alpha, c(3/\sqrt{q}))$-*OWC *protocol for f, where*

$$c\left(1 - \binom{c}{2}\left(\frac{3}{\sqrt{q}}\right)\right) \leq \alpha \leq c.$$

Proof. Let C be the code guaranteed by Theorem 2.1 over alphabet of size q of message length $\beta := \log_q m$, block length $\ell := O_q(\beta)$, and relative distance at least $1 - 3/\sqrt{q}$.

The protocol. Alice on receiving input $x \in \mathcal{X}$ and Bob on receiving input $y \in \mathcal{Y}$ follow the below protocol.

1. Alice and Bob pick a uniformly random $r \in [\ell]$.
2. Alice sends Bob $s := C(x)_r$, i.e., the r^{th} coordinate of the encoding of x.
3. Bob computes the set of field elements, $S := \{C(a)_r\}_{a \in S_y}$, i.e., the r^{th} coordinate of the encoding of all $a \in S_y$.
4. Bob accepts if and only if $s \in S$.

Parameters. It is clear that the above protocol adheres to the structure of an OWC protocol. We now show the specific parameters of the protocol claimed in the theorem statement hold. Alice's message is a field element and thus sends $\lceil \log_2 q \rceil$ bits. The randomness complexity is clearly $\lceil \log_2 \ell \rceil = O_q(1) + \lceil \log_2 \beta \rceil$. Bob accepts only if Alice's message (a field element) is in S, but since f is c-left bounded, we have $|S_y| = c$ and thus $|S| \leq c$. The acceptance complexity of the protocol is clearly the expected size of S over the randomness. Consider the set $C_y = \{C(a) \mid a \in S_y\}$. Since any two codewords of C agree on at most $3/\sqrt{q}$ fraction of coordinates, we have by union bound that there are at least $1 - \binom{c}{2}(3/\sqrt{q})$ fraction of coordinates of $[\ell]$ on which all codewords in C_y are distinct. For such coordinates we have $|S| = c$. Therefore the acceptance complexity is at least $c(1 - \binom{c}{2}(3/\sqrt{q}))$ and at most c.

Completeness. Suppose that $f(x,y) = 1$ then $x \in S_y$ and Bob always accepts.

Soundness. Suppose that $f(x,y) = 0$. This implies that $x \notin S_y$. This implies that for any $a \in S_y$, we have that $C(a)$ and $C(x)$ agree on at most $\ell(3/\sqrt{q})$ coordinates. As before, by taking a union bound we have that there are at most $c\ell(3/\sqrt{q})$ coordinates of $[\ell]$ on which $C(x)$ agrees with $C(a)$ for some $a \in S_y$. Therefore for the remaining coordinates Bob would reject. This implies that Bob rejects with probability at least $1 - c(3/\sqrt{q})$.

By Theorem 2.1, the computation time for Alice and Bob is polynomial time. □

Informally, for large enough q the above theorem gives a $(\log\log m, O(1), c(1 - o(1)), o(1))$-OWC protocol for f. This should be compared with the trivial $(0, \log m, c, 0)$-OWC deterministic protocol for f that was mentioned earlier.

6.2 Connecting OWC protocol to Hardness of Approximating k-median and k-means

Definition 6.3 (Membership function)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be sets of size m and n respectively, where each element in $\mathcal{X}$ is a subset of $\mathcal{Y}$. Then $\mathsf{Mem}_{m,n} : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ is defined by*

$$\mathsf{Mem}_{m,n}(x,y) = \begin{cases} 1 \text{ if } y \in x, \\ 0 \text{ otherwise,} \end{cases} .$$

We rewrite the vertex coverage problem that we had introduced in Section 4 as a special case of the more general max coverage problem. This is done so as to enable us to prove stronger hardness of approximation factors for the ℓ_∞-metric.

Definition 6.4 ((freq, gap)-max coverage)**.** *In the (freq, gap)-max coverage problem, we are given a universe $\mathcal{U}$ of size n, a collection $\mathcal{S}$ of m subsets of $\mathcal{U}$ where each element in $\mathcal{U}$ appears in exactly freq number of subsets in $\mathcal{S}$, and a parameter k as input. We would like to distinguish between the following two cases:*

- **Completeness***: There exists $S_1, \ldots, S_k \in \mathcal{S}$ such that $\bigcup_{i \in [k]} S_i = \mathcal{U}$.*
- **Soundness***: For every $S_1, \ldots, S_k \in \mathcal{S}$ we have $\left| \bigcup_{i \in [k]} S_i \right| \leq \mathsf{gap} \cdot |\mathcal{U}|$.*

We can now rewrite our Theorem 4.2 as follows:

Theorem 6.5 (Austrin and Stanković [AS19])**.** *There is some $\varepsilon > 0$ and $d_0 \in \mathbb{N}$, such that for all $d_{\min} > d_0$, assuming the unique games conjecture, deciding an instance $(\mathcal{U}, \mathcal{S}, k)$ of $(2, 0.9292 - \varepsilon)$-max coverage problem where $|\mathcal{U}| = n$ and $|\mathcal{S}| = poly(n)$, and each set in $\mathcal{S}$ is of cardinality at least $d_{\min}$ is NP-hard.*

We are now ready to state our main connection between OWC-protocols and k-means and k-median inapproximability.

Theorem 6.6. *Let $p \in \mathbb{R}_{\geq 1} \cup \{\infty\}$. Let Π be a (r, μ, α, s)-*OWC *protocol for $\mathsf{Mem}_{m,n}$. Let τ be a λ-gap realization of $H^*(2^\mu, \alpha)$ in the d-dimensional ℓ_p-metric. There is a polynomial time (in input size) algorithm $\mathcal{A}$ which takes as input an instance $(\mathcal{U}, \mathcal{S}, k)$ of the $(\mathsf{freq}, \mathsf{gap})$-max coverage problem where $|\mathcal{U}| = n$ and $|\mathcal{S}| = m$ and outputs an instance $(\mathcal{P}, \mathcal{C}, k)$ of the k-median/k-means problem where we are given a point-set $\mathcal{P} \subset \mathbb{R}^{2^r d}$ of size n, a collection $\mathcal{C}$ of m candidate centers in $\mathbb{R}^{2^r d}$ such that the following holds:*

- **Completeness**: *If there exists $S_1, \ldots, S_k \in \mathcal{S}$ such that $\bigcup_{i \in [k]} S_i = \mathcal{U}$ then there exists $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and $\sigma : \mathcal{P} \to \mathcal{C}'$ such that for all $a \in \mathcal{P}$ we have*

$$\|a - \sigma(a)\|_p = 2^{r/p} \cdot \beta,$$

- **Soundness**: *If for every $S_1, \ldots, S_k \in \mathcal{S}$ we have $\left|\bigcup_{i \in [k]} S_i\right| \leq \mathsf{gap} \cdot |\mathcal{U}|$ then for every $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and every $\sigma : \mathcal{P} \to \mathcal{C}'$ we have that there exists $\mathcal{P}' \subseteq \mathcal{P}$ such that $|\mathcal{P}'| \geq \mathsf{gap} \cdot |\mathcal{P}|$ and for all $a \in \mathcal{P}'$ we have*

$$\|a - \sigma(a)\|_p \geq 2^{r/p} \cdot \lambda \cdot (\alpha + 1 - \mathsf{freq} - s)^{1/p} \cdot \beta,$$

and for all $a \in \mathcal{P} \setminus \mathcal{P}'$ we have

$$\|a - \sigma(a)\|_p = 2^{r/p} \cdot \beta,$$

for some $\beta > 0$.

Proof. Recall from Section 3 that $A^*(t)$ and $B^*(t, r)$ are the partite vertex sets of $H^*(t, r)$. Here we use the short hand $A^* := A^*(2^\mu)$ and $B^* := B^*(2^\mu, \mathsf{freq})$. Also let $\tau : A^* \cup B^* \to \mathbb{R}^d$ and let $\beta > 0$ be the constant from Definition 3.1. We define functions $T_{\mathcal{U}} : \mathcal{U} \times \{0,1\}^r \to B^* \cup \{\perp\}$ and $T_{\mathcal{S}} : \mathcal{S} \times \{0,1\}^r \to A^*$ below. Then, we will construct functions $\widetilde{T}_{\mathcal{U}} : \mathcal{U} \to \mathbb{R}^{d \cdot 2^r}$ and $T_{\mathcal{S}} : \mathcal{S} \to \mathbb{R}^{d \cdot 2^r}$. Given $\widetilde{T}_{\mathcal{U}}$ and $\widetilde{T}_{\mathcal{S}}$ the point-set $\mathcal{P}$ is just defined to be $\{\widetilde{T}_{\mathcal{U}}(u) \mid u \in \mathcal{U}\}$ and the set of candidate centers $\mathcal{C}$ is just $\{\widetilde{T}_{\mathcal{S}}(S) \mid S \in \mathcal{S}\}$.

For every $\gamma \in \{0,1\}^r$ and every $q \in \{0,1\}^\mu$ we define $T_{\mathcal{S}}(S, \gamma) = q$ if in the OWC model where Alice and Bob are trying to compute $\mathsf{Mem}_{m,n} : \mathcal{S} \times \mathcal{U} \to \{0,1\}$, Alice given input S, following the protocol Π would send q on randomness γ. Similarly, we define $R_{u,\gamma} \subseteq \{0,1\}^\mu$, where $q \in \{0,1\}^\mu$ is contained in $R_{u,\gamma}$ if and only if Bob given input u, following the protocol Π would accept the message q sent by Alice on randomness γ. Then, we define $T_{\mathcal{U}}(u, \gamma) = R_{u,\gamma}$ if $|R_{u,\gamma}| = \mathsf{freq}$ and $T_{\mathcal{U}}(u, \gamma) = \perp$ otherwise.

For every possible randomness $\gamma \in \{0,1\}^r$ let c_γ be the number of distinct messages that Bob would accept on input u and randomness γ. Note that $\mathop{\mathbb{E}}_{\gamma \in \{0,1\}^r}[c_\gamma] = \alpha$. Let $\delta := \alpha + 1 - \mathsf{freq}$. From a standard averaging argument, it is easy to see that for every $u \in \mathcal{U}$ there is a subset L_u of $\{0,1\}^r$ of size $2^r \cdot \delta$ such that for all $\gamma \in L_u$ we have $T_{\mathcal{U}}(u, \gamma) \neq \perp$.

Now we can construct functions $\widetilde{T}_{\mathcal{U}} : \mathcal{U} \to \mathbb{R}^{d \cdot 2^r}$ and $\widetilde{T}_{\mathcal{S}} : \mathcal{S} \to \mathbb{R}^{d \cdot 2^r}$ as follows:

$$\forall \gamma \in \{0,1\}^r,\ \widetilde{T}_{\mathcal{U}}(u)|_\gamma = \begin{cases} \tau(T_{\mathcal{U}}(u,\gamma)) \text{ if } T_{\mathcal{U}}(u,\gamma) \neq \perp \\ \tau(\widetilde{u}) \text{ otherwise} \end{cases} \quad \text{and} \quad \widetilde{T}_{\mathcal{S}}(S)|_\gamma = \tau(T_{\mathcal{S}}(S,\gamma)),$$

where $\widetilde{u}$ is any arbitrary element in B^* such that $\widetilde{u}$ is a superset of $R_{u,\gamma}$.

Completeness. Suppose there exist $S_1, \ldots, S_k \in \mathcal{S}$ such that $\bigcup_{i \in [k]} S_i = \mathcal{U}$. Then, we define $\mathcal{C}' = \{\widetilde{T}_{\mathcal{S}}(S_i) \mid i \in [k]\}$. We define $\sigma : \mathcal{P} \to \mathcal{C}'$ as follows: for every $a \in \mathcal{P}$, where $a := \widetilde{T}_{\mathcal{U}}(u)$ for some $u \in \mathcal{U}$, let $\sigma(a)$ be equal to $\widetilde{T}_{\mathcal{S}}(S_i)$ such that $u \in S_i$ (if there is more than one $i \in [k]$ for which S_i contains u then we choose one arbitrarily). Fix $a := \widetilde{T}_{\mathcal{U}}(u)$ in $\mathcal{P}$. Let $c := \sigma(a)$ be the image of S_i under $\widetilde{T}_{\mathcal{S}}$. By definition of σ we have that $u \in S_i$. Therefore, $\mathsf{Mem}_{m,n}(S_i, u) = 1$, and Bob would accept Alice's message for *every* randomness if both of them follow Π.

Fix the randomness γ. Since $\mathsf{Mem}_{m,n}(S_i, u) = 1$ we have that $T_{\mathcal{S}}(S_i, \gamma) \in R_{u,\gamma}$ and thus we have

$$\|\tau(T_{\mathcal{S}}(S_i,\gamma)) - \tau(T_{\mathcal{U}}(u,\gamma))\|_p^p = \beta^p.$$

Summing over all the blocks of coordinates we have:

$$\begin{aligned} \|\widetilde{T}_{\mathcal{U}}(u) - \widetilde{T}_{\mathcal{S}}(S_i)\|_p &= \left(\sum_{\gamma \in L_u} \left(\|\tau(T_{\mathcal{S}}(S_i,\gamma)) - \tau(T_{\mathcal{U}}(u,\gamma))\|_p^p \right) \right)^{1/p} \\ &= 2^{r/p} \cdot \beta \end{aligned} \tag{3}$$

Soundness. Suppose for every $S_1, \ldots, S_k \in \mathcal{S}$ we have $\left| \bigcup_{i \in [k]} S_i \right| \leq \mathsf{gap} \cdot |\mathcal{U}|$. Fix some subset $\mathcal{C}' \subseteq \mathcal{C}$ of size k. Let $\sigma : \mathcal{P} \to \mathcal{C}'$ be some mapping. Consider the mapping $\xi : \mathcal{U} \to \mathcal{S}$ defined by σ as follows. For every $u \in \mathcal{U}$ fix some $S \in \mathcal{S}$ such that $\widetilde{T}_{\mathcal{S}}(S) = \sigma(\widetilde{T}_{\mathcal{U}}(u))$ (in case there are more than one S satisfying $\widetilde{T}_{\mathcal{S}}(S) = \sigma(\widetilde{T}_{\mathcal{U}}(u))$, pick one arbitrarily). Set $\xi(u) = S$. Clearly the range of ξ is of size at most k. Let $\mathcal{S}' = \{S_1, \ldots, S_{k'}\}$ be the range of ξ where $k' \leq k$. We know that there are at least $(1 - \mathsf{gap}) \cdot |\mathcal{U}|$ elements of $\mathcal{U}$ that are not contained in $\bigcup_{i \in [k']} S_i$. Let's call this set $\mathcal{U}'$. Therefore for any $(S, u) \in \mathcal{S}' \times \mathcal{U}'$ we have $\mathsf{Mem}_{m,n}(S, u) = 0$ and Bob would accept Alice's message with probability at most s over the randomness, if both of them follow Π.

Fix some $u \in \mathcal{U}'$. Let $\mathsf{Bad} \subseteq \{0,1\}^r$ such that for all $\gamma \in \mathsf{Bad}$ Bob would reject Alice's message, if both of them follow Π. Similarly, let $\mathsf{Good} \subseteq \{0,1\}^r$ such that for all $\gamma \in \mathsf{Good}$ Bob would accept Alice's message, if both of them follow Π. We have $|\mathsf{Good}| \leq s \cdot 2^r$ and $|\mathsf{Bad}| \geq (1 - s) \cdot 2^r$.

Fix $\gamma \in \mathsf{Bad} \cap L_u$. Since $\mathsf{Mem}_{m,n}(S, u) = 0$ we have that $T_{\mathcal{S}}(S, \gamma) \notin R_{u,\gamma}$ and thus we have

$$\|\tau(T_{\mathcal{S}}(S_i,\gamma)) - \tau(T_{\mathcal{U}}(u,\gamma))\|_p^p \geq \lambda^p \beta^p.$$

Summing over all the blocks of coordinates we have:

$$\begin{aligned}\|\widetilde{T}_{\mathcal{U}}(u) - \widetilde{T}_{\mathcal{S}}(S_i)\|_p &\geq \left(\sum_{\gamma \in \mathsf{Bad} \cap L_u} \left(\|\tau(T_{\mathcal{S}}(S_i,\gamma)) - \tau(T_{\mathcal{U}}(u,\gamma))\|_p^p\right)\right)^{1/p} \\ &= 2^{r/p} \cdot \lambda \cdot \beta \cdot (\delta - s)^{1/p}. \end{aligned} \tag{4}$$

On the other hand, if we fix some $u \in \mathcal{U} \setminus \mathcal{U}'$ then, we have that (3) holds.

Finally, it is easy to see that $\widetilde{T}_{\mathcal{U}}$ and $\widetilde{T}_{\mathcal{S}}$ can be computed in polynomial time as Π is a OWC protocol where the players are bounded to run in $\mathrm{poly}(|\mathcal{U}|, |\mathcal{S}|)$ time. □

7 Inapproximability of k-median and k-means in $O(\log n)$ dimensions

In this section, we finally prove Theorems 1.2, 1.3, and 1.4.

7.1 k-means and k-median in ℓ_p-metric with Candidate Centers

In this subsection, we prove Theorem 1.2.

Theorem 7.1 (k-means with candidate centers in $O(\log n)$ dimensional ℓ_p-metric space)**.** *Let $p \in \mathbb{R}_{\geq 1} \cup \{\infty\}$. Assuming the unique games conjecture, given a point-set $\mathcal{P} \subset \mathbb{R}^d$ of size n (and $d = O(\log n)$), a collection $\mathcal{C}$ of m candidate centers in $\mathbb{R}^d$ (where $m = \mathrm{poly}(n)$), and a parameter k as input, it is* NP*-hard to distinguish between the following two cases:*

- **Completeness***: There exists $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and $\sigma : \mathcal{P} \to \mathcal{C}'$ such that*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_p^2 \leq \beta n \cdot (\log n)^{2/p},$$

- **Soundness***: For every $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and every $\sigma : \mathcal{P} \to \mathcal{C}'$ we have:*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_p^2 \geq \zeta_2(p) \cdot \beta n \cdot (\log n)^{2/p},$$

for some constant $\beta > 0$.

Theorem 7.2 (k-median with candidate centers in $O(\log n)$ dimensional ℓ_p-metric space)**.** *Let $p \in \mathbb{R}_{\geq 1} \cup \{\infty\}$. Assuming the unique games conjecture, given a point-set $\mathcal{P} \subset \mathbb{R}^d$ of size n (and $d = O(\log n)$), a collection $\mathcal{C}$ of m candidate centers in $\mathbb{R}^d$ (where $m = \mathrm{poly}(n)$), and a parameter k as input, it is* NP*-hard to distinguish between the following two cases:*

- **Completeness***: There exists $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and $\sigma : \mathcal{P} \to \mathcal{C}'$ such that*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_p \leq \beta n (\log n)^{1/p},$$

- **Soundness**: *For every $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and every $\sigma : \mathcal{P} \to \mathcal{C}'$ we have:*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_p \geq \zeta_1(p) \cdot \beta n (\log n)^{1/p},$$

for some constant $\beta > 0$.

Proof of Theorems 7.1 and 7.2. At a high level, we use the embedding developed in Proposition 3.9 on the instances built in Theorem 6.6 with the protocol of Theorem 6.2 (setting q to be a super large constant and noting freq to be 2), and then finally use the inapproximability given in Theorem 6.5 to prove the theorems. We provide some additional details below.

Consider the (r, μ, α, s) protocol given in Theorem 6.2. Fix some $\delta > 0$ and $c = 2$. For large enough q we have that $s < \delta$ and $\alpha \in (2 - \delta, 2]$. It is clear that the point-sets $\mathcal{P}$ and candidate centers $\mathcal{C}$ are in dimension $2^{O_q(1)+\log\log m} = O_q(1) \cdot \log m = O_q(\log n)$.

Fix $p \in \mathbb{R}_{\geq 1} \cup \{\infty\}$. Plugging the above calculation into Theorem 6.6, for some fixed constant $\beta > 0$ we have that in the completeness case that the k-median cost in ℓ_p-metric is at most $\beta \cdot n(\log n)^{1/p}$ (and the k-means cost in ℓ_1-metric would be at most $n\beta^2(\log n)^{2/p}$). In the soundness case we have that the k-median cost in ℓ_p-metric is at least

$$\beta \cdot n \cdot (\log n)^{1/p} \cdot (0.0708 \cdot \gamma_p \cdot (1 - 2\delta)^{1/p} + 0.9292) = \beta \cdot n \cdot (\log n)^{1/p} \cdot \zeta_1(p).$$

Similarly from a simple computation of the k-means cost in (4) we have the k-means cost in ℓ_p-metric would be at least

$$\beta \cdot n \cdot (\log n)^{2/p} \cdot (0.0708 \cdot \gamma_p^2 \cdot (1 - 2\delta)^{2/p} + 0.9292) = \beta \cdot n \cdot (\log n)^{2/p} \cdot \zeta_2(p).$$

Note that we can choose δ to be as small as we want. □

7.2 k-median in ℓ_1-metric without Candidate Centers

In this subsection, we prove Theorem 1.4. We also show Theorem 1.3 (now over Boolean vectors).

Theorem 7.3 (k-median in ℓ_1-metric in $O(\log n)$ dimensions without Candidate Centers)**.** *Assuming the unique games conjecture, given a point-set $\mathcal{P} \subset \{0,1\}^d$ of size n (and $d = O(\log n)$) and a parameter k as input, it is* NP*-hard to distinguish between the following two cases:*

- **Completeness**: *There exists $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathbb{R}^d$ and $\sigma : \mathcal{P} \to \mathcal{C}'$ such that*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_2^2 \leq \beta n \log n,$$

- **Soundness**: *For every $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathbb{R}^d$ and every $\sigma : \mathcal{P} \to \mathcal{C}'$ we have:*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_2^2 \geq 1.07 \cdot \beta n \log n,$$

for some constant $\beta > 0$.

Proof. Simply note that after fixing the randomness, the problem on most blocks looks exactly like the instances considered in Theorem 5.2, so the same arguments go through with arbitrarily small loss in approximation factor, in case we restricted our centers to be Boolean valued. Then notice that the translation of the hardness from Hamming metric to ℓ_1-metric as in Theorem 5.4 also holds here. □

Theorem 7.4 (k-means in Euclidean metric in $O(\log n)$ dimensions without Candidate Centers)**.** *Assuming the unique games conjecture, given a point-set $\mathcal{P} \subset \{0,1\}^d$ of size n (and $d = O(\log n)$) and a parameter k as input, it is* NP*-hard to distinguish between the following two cases:*

- **Completeness***: There exists $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathbb{R}^d$ and $\sigma : \mathcal{P} \to \mathcal{C}'$ such that*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_2^2 \leq \beta n \log n,$$

- **Soundness***: For every $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathbb{R}^d$ and every $\sigma : \mathcal{P} \to \mathcal{C}'$ we have:*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_2^2 \geq 1.07 \cdot \beta n \log n,$$

for some constant $\beta > 0$.

Proof. Simply note that after fixing the randomness, the problem on most blocks looks exactly like the instances considered in Theorem 5.5, so the same arguments go through with arbitrarily small loss in approximation factor. □

7.3 Stronger Inapproximability of k-means and k-median in ℓ_∞-metric

In this section, we prove the result for the ℓ_∞-metric given in Theorem 1.1 by a reduction from the gap hypergraph coverage problem. First, we define the gap hypergraph coverage problem.

Let $G(V, E)$ be a hypergraph. Let $S \subseteq V$. We define the cover of S, denoted by $\mathsf{cov}(S)$ as follows:

$$\mathsf{cov}(S) = \{e \in E \mid \exists v \in S \text{ such that } v \in e\}.$$

Definition 7.5 (α-hypergraph coverage)**.** *In the α-hypergraph coverage problem, we are given a hypergraph $G(V, E)$ and a parameter k as input. We would like to distinguish between the following two cases:*

- **Completeness***: There exists $S := \{v_1, \ldots, v_k\} \subseteq V$ such that $\mathsf{cov}(S) = E$.*
- **Soundness***: For every $S := \{v_1, \ldots, v_k\} \subseteq V$ we have $|\mathsf{cov}(S)| \leq \alpha \cdot |E|$.*

We remark here that the hardness of approximation of the minimization version of the hypergraph coverage problem was already shown by Trevisan nearly two decades

ago [Tre01], but standard techniques to convert inapproximability results for optimization problems from minimization instances to maximization instances does not yield the requisite (or even any meaningful) hardness of approximation result that is stated in Theorem 7.6. Nonetheless a careful analysis of the original inapproximability of the max-coverage problem by Feige [Fei98] yields the following.

Theorem 7.6 (Hypergraph Coverage Inapproximability; Essentially [Fei98])**.** *For every $\delta > 0$ there is some $h \in \mathbb{N}$ such that deciding an instance of $(1 - 1/e - \varepsilon)$-hypergraph vertex coverage problem on h-uniform hypergraphs is* NP*-hard.*

We defer the proof of the above theorem to Appendix A. Below are the results in focus of this subsection.

Theorem 7.7 (k-means with candidate centers in $O(\log n)$ dimensional ℓ_∞-metric space)**.** *Let $\varepsilon > 0$. Given a point-set $\mathcal{P} \subset \mathbb{R}^d$ of size n (and $d = O(\log n)$), a collection $\mathcal{C}$ of m candidate centers in $\mathbb{R}^d$ (where $m = poly(n)$), and a parameter k as input, it is* NP*-hard to distinguish between the following two cases:*

- **Completeness**: *There exists $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and $\sigma : \mathcal{P} \to \mathcal{C}'$ such that*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_\infty^2 \leq \beta n,$$

- **Soundness**: *For every $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and every $\sigma : \mathcal{P} \to \mathcal{C}'$ we have:*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_\infty^2 \geq \left(1 + \frac{8}{e} - \varepsilon\right) \cdot \beta n,$$

for some constant $\beta > 0$.

Theorem 7.8 (k-median with candidate centers in $O(\log n)$ dimensional ℓ_∞-metric space)**.** *Let $\varepsilon > 0$. Given a point-set $\mathcal{P} \subset \mathbb{R}^d$ of size n (and $d = O(\log n)$), a collection $\mathcal{C}$ of m candidate centers in $\mathbb{R}^d$ (where $m = poly(n)$), and a parameter k as input, it is* NP*-hard to distinguish between the following two cases:*

- **Completeness**: *There exists $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and $\sigma : \mathcal{P} \to \mathcal{C}'$ such that*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_\infty \leq \beta n,$$

- **Soundness**: *For every $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and every $\sigma : \mathcal{P} \to \mathcal{C}'$ we have:*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_\infty \geq \left(1 + \frac{2}{e} - \varepsilon\right) \cdot \beta n,$$

for some constant $\beta > 0$.

Proof of Theorems 7.7 and 7.8. At a high level, we use the hypergraph embedding developed in Lemma 3.4 on the instances built in Theorem 6.6 with the protocol of Theorem 6.2 (setting q to be a super large constant and noting freq to be some large $h \in \mathbb{N}$), and then finally use the inapproximability given in Theorem 7.6 to prove the theorems. □

8 Open Problems

It remains an important open question to improve upon any of the hardness of approximation results given in Table 1, and in particular to improve upon the inapproximability results of this paper. In this regard, a direction worth pursuing is to design suitable OWC-protcols (for example with acceptance complexity close to 2, arbitrarily small constant soundness, and not too high randomness) for the membership function corresponding to the hypergraph coverage problem. Combining such a protocol with Theorems 6.6 and 7.6 would enable us to prove close to $(1 + 8/e) \approx 3.94$ inapproximability for k-means in the ℓ_1-metric and roughly 1.88 inapproximability for k-means in the Euclidean metric (we refer to both the problems in the discrete case here).

Another interesting open question is to improve upon the $(1 + 8/e)$ inapproximability of k-means or the $(1 + 2/e)$ inapproximability of k-median for any metric space. An obvious barrier to getting such an improvement by starting from the max coverage problem is the triangle inequality.

Finally, we raise the following combinatorial geometry question: can we improve the lower bound given in Lemma 3.6? In particular, can we show that $g_2(H^*(t, 2)) \geq 2$ for large enough t? Notice that when $t = 3$, we have $g_2(H^*(2, 2)) \geq 2$ by placing the six points on the vertices of a regular hexagon in the plane. On the other hand, we suspect that $g_2(H^*(t, 2)) \leq 2$, and confirming such a claim would also be interesting.

Acknowledgements

We are truly grateful to Per Austrin, Amey Bhangale, Euiwoong Lee, and Pasin Manurangsi for discussions about the inapproximability of the vertex coverage and max coverage problem.

References

[ACKS15] Pranjal Awasthi, Moses Charikar, Ravishankar Krishnaswamy, and Ali Kemal Sinop. The hardness of approximation of euclidean k-means. In *31st International Symposium on Computational Geometry, SoCG 2015, June 22-25, 2015, Eindhoven, The Netherlands*, pages 754–767, 2015.

[AKS11] Per Austrin, Subhash Khot, and Muli Safra. Inapproximability of vertex cover and independent set in bounded degree graphs. *Theory of Computing*, 7(1):27–43, 2011.

[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.

[ANSW16] Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, and Justin Ward. Better guarantees for k-means and euclidean k-median by primal-dual algorithms. *CoRR*, abs/1612.07925, 2016.

[ARW17] Amir Abboud, Aviad Rubinstein, and R. Ryan Williams. Distributed PCP theorems for hardness of approximation in P. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 25–36, 2017.

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

[AS19] Per Austrin and Aleksa Stanković. Global cardinality constraints makes approximating some max-2-csps harder. In *APPROX*, 2019.

[AV07] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007*, pages 1027–1035, 2007.

[BGS98] Mihir Bellare, Oded Goldreich, and Madhu Sudan. Free bits, pcps, and nonapproximability-towards tight results. *SIAM J. Comput.*, 27(3):804–915, 1998.

[BK19] Amey Bhangale and Subhash Khot. Ug-hardness to np-hardness by losing half. *Electronic Colloquium on Computational Complexity (ECCC)*, 26:4, 2019.

[BKS19] Boaz Barak, Pravesh K. Kothari, and David Steurer. Small-set expansion in shortcode graph and the 2-to-2 conjecture. In *10th Innovations in Theoretical Computer Science Conference, ITCS 2019, January 10-12, 2019, San Diego, California, USA*, pages 9:1–9:12, 2019.

[BPR+15] Jaroslaw Byrka, Thomas Pensyl, Bartosz Rybicki, Aravind Srinivasan, and Khoa Trinh. An improved approximation for *k*-median, and positive correlation in budgeted optimization. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, pages 737–756, 2015.

[CADMRR18] Vincent Cohen-Addad, Arnaud De Mesmay, Eva Rotenberg, and Alan Roytman. The bane of low-dimensionality clustering. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 441–456. SIAM, 2018.

[CAGK+19] Vincent Cohen-Addad, Anupam Gupta, Amit Kumar, Euiwoong Lee, and Jason Li. Tight FPT Approximations for k-Median and k-Means. In Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, and Stefano Leonardi, editors, *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*, volume 132 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 42:1–42:14, Dagstuhl, Germany, 2019. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[CAKM16] Vincent Cohen-Addad, Philip N. Klein, and Claire Mathieu. Local search yields approximation schemes for k-means and k-median in euclidean

and minor-free metrics. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 353–364, 2016.

[CFS18] Vincent Cohen-Addad, Andreas Emil Feldmann, and David Saulpic. Near-linear time approximation schemes for clustering in doubling metrics. *CoRR*, abs/1812.08664, 2018.

[Che18] Lijie Chen. On the hardness of approximate and exact (bichromatic) maximum inner product. In *33rd Computational Complexity Conference, CCC 2018, June 22-24, 2018, San Diego, CA, USA*, pages 14:1–14:45, 2018.

[Coh18] Vincent Cohen-Addad. A fast approximation scheme for low-dimensional k-means. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*, pages 430–440, 2018.

[Das08] Sanjoy Dasgupta. *The hardness of k-means clustering*. Technical Report, Department of Computer Science and Engineering, University of California, San Diego, 2008.

[DKK+18a] Irit Dinur, Subhash Khot, Guy Kindler, Dor Minzer, and Muli Safra. On non-optimally expanding sets in grassmann graphs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 940–951, 2018.

[DKK+18b] Irit Dinur, Subhash Khot, Guy Kindler, Dor Minzer, and Muli Safra. Towards a proof of the 2-to-1 games conjecture? In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 376–389, 2018.

[DKL19] Roee David, Karthik C. S., and Bundit Laekhanukit. On the complexity of closest pair via polar–pair of point–sets. *SIAM J. Discrete Math.*, 33(1):509–527, 2019.

[dlVKKR03] Wenceslas Fernandez de la Vega, Marek Karpinski, Claire Kenyon, and Yuval Rabani. Approximation schemes for clustering problems. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing, June 9-11, 2003, San Diego, CA, USA*, pages 50–58, 2003.

[DS14] Irit Dinur and David Steurer. Analytical approach to parallel repetition. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 624–633, 2014.

[EIO02] Lars Engebretsen, Piotr Indyk, and Ryan O'Donnell. Derandomized dimensionality reduction with applications. In *Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms, January 6-8, 2002, San Francisco, CA, USA.*, pages 705–712, 2002.

[Fei98] Uriel Feige. A threshold of ln n for approximating set cover. *J. ACM*, 45(4):634–652, 1998.

[FM86] Peter Frankl and Hiroshi Maehara. Embedding the n-cube in lower dimensions. *Eur. J. Comb.*, 7(3):221–225, 1986.

[FM88] Peter Frankl and Hiroshi Maehara. On the contact dimensions of graphs. *Discrete & Computational Geometry*, 3:89–96, 1988.

[FRS16] Zachary Friggstad, Mohsen Rezapour, and Mohammad R. Salavatipour. Local search yields a PTAS for k-means in doubling metrics. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 365–374, 2016.

[GI03] Venkatesan Guruswami and Piotr Indyk. Embeddings and non-approximability of geometric problems. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, January 12-14, 2003, Baltimore, Maryland, USA.*, pages 537–538, 2003.

[Gil52] E. N. Gilbert. A comparison of signalling alphabets. *Bell System Technical Journal*, 31:504 – 522, 1952.

[GK99] Sudipto Guha and Samir Khuller. Greedy strikes back: Improved facility location algorithms. *J. Algorithms*, 31(1):228–248, 1999.

[GS96] Arnaldo Garcia and Henning Stichtenoth. On the asymptotic behaviour of some towers of function fields over finite fields. *Journal of Number Theory*, 61(2):248 – 273, 1996.

[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.

[KLM18] Karthik C. S., Bundit Laekhanukit, and Pasin Manurangsi. On the parameterized complexity of approximating dominating set. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 1283–1296, 2018.

[KM19] Karthik C. S. and Pasin Manurangsi. On closest pair in euclidean metric: Monochromatic is as hard as bichromatic. In *10th Innovations in Theoretical Computer Science Conference, ITCS 2019, January 10-12, 2019, San Diego, California, USA*, pages 17:1–17:16, 2019.

[KMS17] Subhash Khot, Dor Minzer, and Muli Safra. On independent sets, 2-to-2 games, and grassmann graphs. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 576–589, 2017.

[KMS18] Subhash Khot, Dor Minzer, and Muli Safra. Pseudorandom sets in grassmann graph have near-perfect expansion. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pages 592–601, 2018.

[KN97] Eyal Kushilevitz and Noam Nisan. *Communication Complexity*. Cambridge University Press, New York, NY, USA, 1997.

[KR07] Stavros G. Kolliopoulos and Satish Rao. A nearly linear-time approximation scheme for the euclidean k-median problem. *SIAM J. Comput.*, 37(3):757–782, June 2007.

[KSS04] Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time (1+ ") -approximation algorithm for k-means clustering in any dimensions. In *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '04, pages 454–462, Washington, DC, USA, 2004. IEEE Computer Society.

[LSW17] Euiwoong Lee, Melanie Schmidt, and John Wright. Improved and simplified inapproximability for k-means. *Inf. Process. Lett.*, 120:40–43, 2017.

[LY94] Carsten Lund and Mihalis Yannakakis. On the hardness of approximating minimization problems. *J. ACM*, 41(5):960–981, 1994.

[Mae85] Hiroshi Maehara. Contact patterns of equal nonoverlapping spheres. *Graphs and Combinatorics*, 1(1):271–282, 1985.

[Mae91] Hiroshi Maehara. Dispersed points and geometric embedding of complete bipartite graphs. *Discrete & Computational Geometry*, 6:57–67, 1991.

[Man19] Pasin Manurangsi. A note on max k-vertex cover: Faster fpt-as, smaller approximate kernel and improved approximation. In *2nd Symposium on Simplicity in Algorithms, SOSA@SODA 2019, January 8-9, 2019 - San Diego, CA, USA*, pages 15:1–15:21, 2019.

[Mos15] Dana Moshkovitz. The projection games conjecture and the np-hardness of ln n-approximating set-cover. *Theory of Computing*, 11:221–235, 2015.

[MS84] Nimrod Megiddo and Kenneth J. Supowit. On the complexity of some common geometric location problems. *SIAM J. Comput.*, 13(1):182–196, 1984.

[Pac80] Janos Pach. Decomposition of multiple packing and covering. *Diskrete Geometrie*, 2 Kolloq. Math. Inst. Univ. Salzburg:169–178, 1980.

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

[Rub18] Aviad Rubinstein. Hardness of approximate nearest neighbor search. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 1260–1268, 2018.

[SAK+01] Kenneth W. Shum, Ilia Aleshnikov, P. Vijay Kumar, Henning Stichtenoth, and Vinay Deolalikar. A low-complexity algorithm for the construction of algebraic-geometric codes better than the Gilbert-Varshamov bound. *IEEE Trans. Information Theory*, 47(6):2225–2241, 2001.

[Tre00] Luca Trevisan. When hamming meets euclid: The approximability of geometric TSP and steiner tree. *SIAM J. Comput.*, 30(2):475–485, 2000.

[Tre01] Luca Trevisan. Non-approximability results for optimization problems on bounded degree instances. In *Proceedings on 33rd Annual ACM Symposium on Theory of Computing, July 6-8, 2001, Heraklion, Crete, Greece*, pages 453–461, 2001.

[Var57] R. R. Varshamov. Estimate of the number of signals in error correcting codes. *Dokl. Akad. Nauk SSSR*, 117:739 – 741, 1957.

[Yao79] Andrew Chi-Chih Yao. Some complexity questions related to distributive computing (preliminary report). In *STOC*, pages 209–213, 1979.

A Inapproximability of Hypergraph Vertex Coverage

In this section, we show Theorem 7.6 which essentially follows from Feige's proof [Fei98] of the hardness of approximation of max-coverage problem. However, we present the proof below in terms of label cover (as in [Mos15, DS14]) instead of multi-prover proof systems (as in [LY94, Fei98]).

Before we delve into the proof of Theorem 7.6, we formally define the label cover problem and state its hardness of approximation result that follows from the application of the parallel repetition theorem [Raz98, DS14] to the PCP theorem [AS98, ALM+98]. Below we state a restricted bounded degree and bounded alphabet size version of gap label cover problem.

Definition A.1 (Label Cover problem[9])**.** *Let $\varepsilon > 0$, $d, \alpha \in \mathbb{N}$. Let Σ_U, Σ_V be two finite sets. The input to a (ε, d, α)-label cover problem Π is a bipartite graph $G(U \cup V, E)$ and a set of projection functions $\pi = \{\pi_e : \Sigma_U \to \Sigma_V \mid e \in E\}$ such that the following holds:*

- $|\Sigma_U|, |\Sigma_V| \leq \alpha$.
- *for all $u \in U \cup V$, we have degree of u is at most d.*

For every assignment $\sigma := (\sigma_U : U \to \Sigma_U, \sigma_V : V \to \Sigma_V)$ to Π, we define $\mathsf{sat}(\Pi, \sigma)$ as follows:

$$\mathsf{sat}(\Pi, \sigma) := \mathop{\mathbb{E}}_{e:=(u,v)\sim E}[\pi_e(\sigma_U(u)) = \sigma_V(v)].$$

The goal of the (ε, d, α)-label cover problem is to distinguish between the following two cases.

- **Completeness**: *There exists an assignment σ to Π such that $\mathsf{sat}(\Pi, \sigma) = 1$.*
- **Soundness**: *For every assignment σ to Π we have that $\mathsf{sat}(\Pi, \sigma) \leq \varepsilon$.*

[9] The label cover problem as defined here is known in literature as the label cover problem with projection property or as the projection game problem, but we drop the word 'projection' here for brevity.

An immediate consequence of the PCP theorem is that it is NP-hard to decide an instance $\Pi(G, \pi)$ of (ε, d, α)-label cover problem for some constants $\varepsilon > 0, d, \alpha \in \mathbb{N}$. By applying the parallel repetition theorem to the gap instances arising from the PCP theorem, we get the following.

Theorem A.2 (Bounded Label Cover Inapproximability [AS98, ALM+98, Raz98])**.** *For every constant $\varepsilon > 0$, there exist constants $d := d(\varepsilon) \in \mathbb{N}$ and $\alpha := \alpha(\varepsilon) \in \mathbb{N}$ such that it is* NP*-hard to decide an instance $\Pi(G, \pi)$ of (ε, d, α)-label cover problem.*

Before we proceed to the hardness of approximation of Hypergraph Vertex Coverage problem, we do the following preprocessing step on the label cover instances.

Theorem A.3 (Label Cover Inapproximability with total disagreement [Mos15])**.** *For every constant $\varepsilon > 0$, there exist constants $d := d(\varepsilon) \in \mathbb{N}$ and $\alpha := \alpha(\varepsilon) \in \mathbb{N}$ such that given as input a instance $\Pi(G, \pi)$ of (ε, d, α)-label cover problem, it is* NP*-hard to distinguish between the following two cases.*

- **Completeness***: There exists an assignment σ to Π such that $\mathsf{sat}(\Pi, \sigma) = 1$.*
- **Soundness***: For every assignment σ to Π we have that $\mathsf{sat}(\Pi, \sigma) \leq \varepsilon$.*

Proof. □

Theorem A.4 (Essentially Feige [Fei98])**.** *For every $\delta > 0$ there is some $h \in \mathbb{N}$ such that deciding an instance of $(1 - 1/e - \delta)$-hypergraph vertex coverage problem on hypergraphs of arity at most h is* NP*-hard.*

Proof. Fix $\delta > 0$. We define $\varepsilon :=$. We reduce an instance $\Pi(G(U \cup V, E), \pi)$ of (ε, d, α)-label cover problem to an instance of $(1 - 1/e - \delta)$-hypergraph vertex coverage problem on hypergraphs of arity at most h, where d and α are constants depending only on ε as set by Theorem A.4 and $h :=$. The theorem statement then follows from Theorem A.4.

We now build a hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ which is an instance of the $(1 - 1/e - \delta)$-hypergraph vertex coverage problem. We define $\mathcal{V} := U \times \Sigma_U$. For every $v \in V$, we build a set of hyperedges $\mathcal{E}_v$ over $\mathcal{V}$ as follows:

$$\mathcal{E}_v = \{\}.$$

□

say universe element

B Inapproximability of Clustering in Edit Metric

In this section, we show how our inapproximability results for k-median and k-means can be extended to the edit metric. First, we recall the following technical tool established in [Rub18].

Lemma B.1 (Rubinstein [Rub18])**.** *For large enough $d \in \mathbb{N}$, there is a function $\eta : \{0,1\}^d \to \{0,1\}^{d'}$, where $d' = O(d \log d)$, such that for all $a, b \in \{0,1\}^d$ the following holds for some constant $\lambda > 0$:*

$$|\mathsf{ed}(\eta(a), \eta(b)) - \lambda \cdot \log d \cdot \|a - b\|_0| = o(d').$$

Moreover, for any $a \in \{0,1\}^d$, $\eta(a)$ can be computed in $2^{o(d)}$ time.

We state our hardness of approximation result for the edit metric below.

Theorem B.2 (k-means with candidate centers in $O(\log n \cdot \log\log n)$ dimensional Edit-metric space)**.** *Assuming the unique games conjecture, given a point-set $\mathcal{P} \subset \{0,1\}^d$ of size n (and $d = O(\log n \log\log n)$), a collection $\mathcal{C}$ of m candidate centers in $\{0,1\}^d$ (where $m = poly(n)$), and a parameter k as input, it is* NP*-hard to distinguish between the following two cases:*

- **Completeness**: *There exists $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and $\sigma : \mathcal{P} \to \mathcal{C}'$ such that*

$$\sum_{a \in \mathcal{P}} \mathsf{ed}(a, \sigma(a))^2 \leq n \cdot \beta(n),$$

- **Soundness**: *For every $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and every $\sigma : \mathcal{P} \to \mathcal{C}'$ we have:*

$$\sum_{a \in \mathcal{P}} \mathsf{ed}(a, \sigma(a))^2 \geq 1.56 \cdot n \cdot \beta(n),$$

for some fixed $\beta : \mathbb{N} \to \mathbb{N}$ such that $\beta(n) = polylog(n)$.

Theorem B.3 (k-median with candidate centers in $O(\log n \cdot \log\log n)$ dimensional Edit-metric space)**.** *Assuming the unique games conjecture, given a point-set $\mathcal{P} \subset \{0,1\}^d$ of size n (and $d = O(\log n \log\log n)$), a collection $\mathcal{C}$ of m candidate centers in $\{0,1\}^d$ (where $m = poly(n)$), and a parameter k as input, it is* NP*-hard to distinguish between the following two cases:*

- **Completeness**: *There exists $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and $\sigma : \mathcal{P} \to \mathcal{C}'$ such that*

$$\sum_{a \in \mathcal{P}} \mathsf{ed}(a, \sigma(a)) \leq n \cdot \beta(n),$$

- **Soundness**: *For every $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and every $\sigma : \mathcal{P} \to \mathcal{C}'$ we have:*

$$\sum_{a \in \mathcal{P}} \mathsf{ed}(a, \sigma(a)) \geq 1.14 \cdot n \cdot \beta(n),$$

for some fixed $\beta : \mathbb{N} \to \mathbb{N}$ such that $\beta(n) = polylog(n)$.

Proof of Theorems B.2 and B.3. The proof follows from the hard instances constructed in the proof of Theorem 7.1 for the Hamming metric. More precisely, in the proof of Theorem 7.1, we reduce to the following problem: given a point-set $\mathcal{P} \subset \{0,1\}^d$ of size n (and $d = O(\log n)$), a collection $\mathcal{C}$ of m candidate centers in $\{0,1\}^d$ (where $m = \text{poly}(n)$), and a parameter k as input, it is NP-hard to distinguish between the following two cases:

- **Completeness**: There exists $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and $\sigma : \mathcal{P} \to \mathcal{C}'$ such that for all $a \in \mathcal{P}$, we have

$$\|a - \sigma(a)\|_0 = \beta \log n, \tag{5}$$

- **Soundness**: For every $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and every $\sigma : \mathcal{P} \to \mathcal{C}'$ we have that there exists $\mathcal{P}_\sigma \subseteq \mathcal{P}$ such that $|\mathcal{P}_\sigma| \geq (0.0708 + \varepsilon) \cdot |\mathcal{P}|$, for some small $\varepsilon > 0$, we have

$$\forall a \in \mathcal{P}_\sigma,\ \|a - \sigma(a)\|_0 \geq (3 - \delta) \cdot \beta \log n, \tag{6}$$
$$\forall a \in \mathcal{P} \setminus \mathcal{P}_\sigma,\ \|a - \sigma(a)\|_0 = \beta \log n, \tag{7}$$

for some constant $\beta > 0$ and any $\delta > 0$.

We consider the above given point-set $\mathcal{P} \subset \{0,1\}^d$ and the collection $\mathcal{C}$ of candidate centers in $\{0,1\}^d$, and construct the input point-set $\mathcal{P}^* \subset \{0,1\}^{d'}$ and the collection $\mathcal{C}^*$ of candidate centers in $\{0,1\}^{d'}$ for the edit-metric, where $d' = O(d \log d)$. We define $\mathcal{P}^*$ and $\mathcal{C}^*$ as follows:

$$\mathcal{P}^* = \{\eta(a) \mid a \in \mathcal{P}\},\ \ \mathcal{C}^* = \{\eta(c) \mid c \in \mathcal{C}\},$$

where η was as given in Lemma B.1.

Let us suppose that (5) holds. Consider $\mathcal{C}'' \subseteq \mathcal{C}^*$ defined as

$$\mathcal{C}'' := \{\eta(c_i) | c_i \in \mathcal{C}'\}.$$

Then, we have that for all $a^* \in \mathcal{P}^*$,

$$\begin{aligned}\mathsf{ed}(a^*, \eta(\sigma(\eta^{-1}(a^*)))) &\leq \lambda \cdot \log d \cdot \|a - \sigma(a)\|_0 + o(d') \\ &= \beta \cdot \lambda \cdot \log n \cdot \log\log n \cdot (1 + o(1)).\end{aligned}$$

Now, let us suppose that (6) and (7) holds. Consider any $\mathcal{C}'' \subseteq \mathcal{C}^*$ such that $|\mathcal{C}''| = k$ and any $\sigma' : \mathcal{P}^* \to \mathcal{C}''$. We now define $\mathcal{C}' \subseteq \mathcal{C}$ and $\sigma : \mathcal{P} \to \mathcal{C}'$ as follows

$$\mathcal{C}' := \{\eta^{-1}(c) | c \in \mathcal{C}''\}, \quad \text{and } \forall a \in \mathcal{P}, \sigma(a) = \sigma'(\eta(a)).$$

Define $\mathcal{P}^*_{\sigma'} := \{\eta(a) \mid a \in \mathcal{P}_\sigma\}$. Then, we have,

$$\begin{aligned}\forall a^* \in \mathcal{P}^*_{\sigma'},\ \mathsf{ed}(a^*, \sigma'(a^*)) &\geq \lambda \cdot \log d \cdot \|\eta^{-1}(a^*) - \sigma(\eta^{-1}(a^*))\|_0 - o(d'), \\ &\geq \beta \cdot \lambda \cdot \log n \cdot \log\log n \cdot (3 - \delta - o(1)), \\ \forall a \in \mathcal{P}^* \setminus \mathcal{P}^*_{\sigma'},\ \mathsf{ed}(a^*, \sigma'(a^*)) &\geq \lambda \cdot \log d \cdot \|\eta^{-1}(a^*) - \sigma(\eta^{-1}(a^*))\|_0 - o(d') \\ &= \beta \cdot \lambda \cdot \log n \cdot \log\log n \cdot (1 - o(1)).\end{aligned}$$

The proof of the theorem statements then follows by noting that $|\mathcal{P}^*_{\sigma'}| = |\mathcal{P}_\sigma| \geq (0.0708 + \varepsilon) \cdot |\mathcal{P}| \geq (0.0708 + \varepsilon) \cdot |\mathcal{P}^*|$. □