Electronic Colloquium on Computational Complexity, Report No. 177 (2021)

Johnson Coverage Hypothesis: Inapproximability of k-means and k-median in ℓ_p-metrics

Vincent Cohen-Addad* Karthik C. S.† Euiwoong Lee‡

November 21, 2021

Abstract

k-median and k-means are the two most popular objectives for clustering algorithms. Despite intensive effort, a good understanding of the approximability of these objectives, particularly in ℓ_p-metrics, remains a major open problem. In this paper, we significantly improve upon the hardness of approximation factors known in literature for these objectives in ℓ_p-metrics.

We introduce a new hypothesis called the *Johnson Coverage Hypothesis* (JCH), which roughly asserts that the well-studied Max k-Coverage problem on set systems is hard to approximate to a factor greater than $(1 - 1/e)$, even when the membership graph of the set system is a subgraph of the Johnson graph. We then show that together with generalizations of the embedding techniques introduced by Cohen-Addad and Karthik (FOCS '19), JCH implies hardness of approximation results for k-median and k-means in ℓ_p-metrics for factors which are close to the ones obtained for general metrics. In particular, assuming JCH we show that it is hard to approximate the k-means objective:

- Discrete case: To a factor of 3.94 in the ℓ_1-metric and to a factor of 1.73 in the ℓ_2-metric; this improves upon the previous factor of 1.56 and 1.17 respectively, obtained under the Unique Games Conjecture (UGC).
- Continuous case: To a factor of 2.10 in the ℓ_1-metric and to a factor of 1.36 in the ℓ_2-metric; this improves upon the previous factor of 1.07 in the ℓ_2-metric obtained under UGC (and to the best of our knowledge, the continuous case of k-means in ℓ_1-metric was not previously analyzed in literature).

We also obtain similar improvements under JCH for the k-median objective.

Additionally, we prove a weak version of JCH using the work of Dinur et al. (SICOMP '05) on Hypergraph Vertex Cover, and recover all the results stated above of Cohen-Addad and Karthik (FOCS '19) to (nearly) the same inapproximability factors but now under the standard NP $\neq$ P assumption (instead of UGC).

Finally, we establish a strong connection between JCH and the long standing open problem of determining the Hypergraph Turán number. We then use this connection to prove improved SDP gaps (over the existing factors in literature) for k-means and k-median objectives.

*Google Research, Switzerland. vcohenad@gmail.com.
†Rutgers University, USA. karthik.cs@rutgers.edu.
‡University of Michigan, USA. euiwoong@umich.edu.

ISSN 1433-8092

1 Introduction

Over the last 15 years, the Unique Game Conjecture has enabled tremendous progress on the understanding of the (in)approximability of fundamental graph problems, such as Vertex Cover [KR08], Max Cut [KKMO07], Betweeness [CGM09], and Sparsest Cut [CKK+06, KV15, AKK+08], sometimes even leading to tight results. Yet, the approximability of most NP-hard *geometric* problems remains wide open: even for fundamental geometric problems such as the Traveling Salesman Problem, Steiner Tree, k-median, and Facility Location, the hardness of approximation in various ℓ_p-metrics has remained below 1.01. On the other hand, approximation algorithms for the aforementioned problems in $\Omega(\log n)$ dimensions do not achieve much better guarantees than what they do for general metrics. This situation stands in stark contrast to the bounded dimension versions of the same problems, whose approximabilities are mostly well understood.

The main approach for proving hardness of approximation for geometric problems in ℓ_p-metrics consists of two components: (1) establishing hardness of approximation in general metric spaces; (2) finding an embedding of the hard instances into ℓ_p-metrics that preserves the gap. The challenge here is to make (1) and (2) *meet* at a sweet spot: we would like the hard gap instances in general metric spaces to be 'embeddable', i.e., to be mapped into the ℓ_p-metric space while preserving the distances between points, but, proving a significant inapproximability bound for 'embeddable' instances requires a deep understanding of the hard instances (in the general metric) of the problem at hand. This paper aims at characterizing the sweet spot between (1) and (2), hence providing a general framework for obtaining strong inapproximability for a large family of clustering and covering problems.

Hardness of Clustering Problems in ℓ_p-metrics. Given a set of points in a metric space, a clustering is a partition of the points such that points in the same part are close to each other. Thus studying the complexity of finding good clustering is a very natural research avenue: On one hand these problems have a wide range of applications, ranging from unsupervised learning, to information retrieval, and bioinformatics; on the other hand, as we will illustrate in this paper, such problems are very natural generalizations of set-cover-type problems to the metric setting, and as such are very fundamental computational problems. The most popular objectives for clustering in metric spaces are arguably the k-median and k-means problems: Given a set of points P in a metric space, the k-median problem asks to identify a set of k representatives, called *centers*, such that the sum of the distances from each point to its closest center is minimized (for the k-means problem, the goal is to minimize the sum of distances squared) – see Section 2 for formal definitions. In general metrics, the k-median and k-means problems are known to be hard to approximate within a factor 1.73 and 3.94 respectively [GK99], whereas the best known approximation algorithms achieve an approximation guarantee of 2.67 and 9 respectively [BPR+15, ANSW20].

A natural question arising from the above works is whether one can exploit the structure of more specific metrics, such as doubling or Euclidean metrics to obtain better approximation or bypass the lower bound. If the points lie in an Euclidean space of arbitary dimension, a 6.357-approximation and a 2.633-approximation are known for k-means and k-median respectively [ANSW20], while a near-linear time approximation scheme is known for doubling metrics [CSF19][1]. In terms of hardness of approximation: the problems were known to be APX-Hard since the early 2000s [Tre00, GI03] in Euclidean spaces of dimension $\Omega(\log n)$ and have

[1] With doubly exponential dependency in the dimension

recently been shown to be hard to approximate within a factor of 1.17 and 1.06 for k-means and k-median respectively [CK19]. Perhaps surprisingly, for some other structured metrics such as Hamming or Edit distance, no better approximation algorithms are known while the hardness known for both Hamming and Edit distances were 1.56 for k-means problem and 1.14 for k-median problem. Summarizing, the gap between upper and lower bounds for the Euclidean, Hamming, and Edit metrics, remains huge.

Technical Barriers. A well-known framework to obtain hardness of approximation results in the general metric for clustering objectives is through a straightforward reduction from the Max k-Coverage[2] or the Set Cover problem. We create a 'point' for each element of the universe and a 'candidate center', namely a location where it is possible to place a center, for each set. Then, we define the distance between a point (corresponding to an element of the universe) and a candidate center (corresponding to a set) to be 1 if the set contains the element and 3 otherwise. This reduction due to Guha and Khuller [GK99] yields lower bounds of $1 + 2/e$ and $1 + 8/e$ for the k-median and k-means problems, respectively, in general discrete metric spaces.

However, for the k-median and k-means objectives, the above strong hardness results for Max k-Coverage produce instances that are impossible to embed in $\mathbb{R}^n$ without suffering a huge distortion in the distances, and thus would yield only a trivial gap for k-means or k-median problems in ℓ_p-metrics. This issue has been faced for most other geometric problems as well, such as TSP or Steiner tree. It is tempting to obtain more structure on the "hard instances" of set-cover-type problems with the large toolbox that has been developed around the Unique Games Conjecture over the last 15 years for fundamental graph problems such as Vertex Cover or Sparsest Cut. However, it seems that most of these hardness of approximation tools would not provide the adequate structure on the set cover instances constructed, namely a structure that would make the whole instance easily "embeddable" into an ℓ_p-metric.

This illustrates the main difficulty in understanding k-median / k-means in ℓ_p-metrics; Guha-Khuller type reductions from general set systems with strong (perhaps optimal) gaps are not directly "embeddable", but the current results use hardness for very restricted systems in a black-box way and thus cannot be extended to give strong hardness of approximation results. Thus, we ask:

What structure on hard instances of Max k-Coverage problem would yield meaningful hardness of k-median and k-means in ℓ_p-metrics?

1.1 Our Results

This paper addresses the above question in a unified manner, providing a general framework for obtaining strong inapproximability in ℓ_p-metrics (see Table 1 for a summary of our results). We segregate our results below in terms of conceptual and technical contributions and provide more details in the subsequent subsubsections.

Our main conceptual contribution is proposing the Johnson Coverage Hypothesis (JCH) and identifying it as lying at the heart of the hard instances of k-means and k-median problems in ℓ_p-metrics. Intuitively, JCH conjectures that the $(1 - 1/e)$-hardness of approximation for Max k-Coverage [Fei98] holds even for set systems whose bipartite incidence graph (between

[2]Given a set system and $k \in \mathbb{N}$, the problem asks to choose k sets to maximize the size of their union.

Metric	Discrete k-means	Discrete k-median	Continuous k-means	Continuous k-median	Assumption	
ℓ_0	1.56 [CK19]	1.14 [CK19]	1.21 [CK19]	1.07 [CK19]	UGC	Previous
	1.38	**1.12**	**1.16**	**1.06**	NP $\neq$ P	This paper
	3.94	**1.73**	**2.10**	**1.36**	JCH/JCH*	This paper
ℓ_1	1.56 [CK19]	1.14 [CK19]	–	1.07 [CK19]	UGC	Previous
	1.38	**1.12**	**1.16**	**1.06**	NP $\neq$ P	This paper
	3.94	**1.73**	**2.10**	**1.36**	JCH/JCH*	This paper
ℓ_2	1.17 [CK19]	1.06 [CK19]	1.07 [CK19]	–	UGC	Previous
	1.17	**1.07**	**1.06**	**1.015**	NP $\neq$ P	This paper
	1.73	**1.27**	**1.36**	**1.08**	JCH/JCH*	This paper
ℓ_∞	3.94 [CK19]	1.73 [CK19]	Open[3]	Open[3]	NP $\neq$ P	Previous
General	3.94 [GK99]	1.73 [GK99]	4 [CKL21]	2 [CKL21]	NP $\neq$ P	Previous

Table 1: 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. All the results for ℓ_p-metrics stated inside the table are when the input points are given in $O(\log n)$ dimensions. The NP-hardness reductions for the continuous objectives are randomized.

sets and elements) is a subgraph of the *Johnson graph* (see Section 1.1.1 for the formal definition). Johnson graphs are well-studied objects in combinatorics that also admit nice geometric embedding properties. More generally, we argue that identifying an intermediate mathematical property between geometry and combinatorics, and revisiting fundamental combinatorial optimization problems (Max k-Coverage in this paper) restricted to instances having that property is a fruitful avenue to understand approximability of high dimensional geometric optimization problems.

Our main technical contributions are two-fold. First, we generalize the embedding technique introduced in [CK19] that gives a reduction from covering problems to clustering problems. The embedding technique in [CK19] has two components: the first component is embedding the incidence graph of the complete graph into ℓ_p-metrics and the second component is a dimension reduction technique developed by designing efficient protocols for the Vertex Cover problem in the communication model. We generalize both these components: we show how to embed the incidence graph of the complete *hypergraph* into ℓ_p-metrics and we provide a dimension reduction technique by designing efficient protocols for the *Set Cover* problem in the communication model. These generalization were necessary to prove strong inapproximability results for clustering objectives as we needed to handle the less structured instances arising from JCH and its variants (as opposed to [CK19] whose reduction started from the more structured Vertex Coverage problem).

Combining JCH with the new embedding results above, we deduce strong and perhaps even surprising inapproximability results: for example, we show that k-means and k-median problems are no easier in the ℓ_1-metric/Hamming metric/Edit distance metric than in general

[3]See Open Problem 5.5.

metric spaces[4], or obtain a much higher inapproximability for ℓ_2-metric (see Table 1 and Section 1.1.3 for the whole list of results we obtain). Indeed, our framework accommodates hardness of *any* variant of JCH defined in Section 1.1.2 in order to show new hardness of clustering problems; for instance, the classic Set Cover hardness of Feige (or any APX-hardness of Max k-Coverage) can be plugged into our framework to produce APX-Hardness for k-means and k-median in all ℓ_p-metrics in both the discrete and the continuous case (see Remarks 3.14 and 3.19 for precise details)

Our second technical contribution is a new NP-Hardness of approximation result for a (weaker) variant of JCH corresponding to Hypergraph Vertex Coverage in 3-uniform hypergraphs. Consequently, with the above generalized embedding technique, we obtained NP-Hardness results for clustering problems to factors that nearly match and sometimes even improve upon the previous best hardness results (which were based on the Unique Games Conjecture); see Table 1 for the precise factors. Our proof relies on (i) tighter analysis of the *Multilayered PCP* constructed by Dinur et al. [DGKR05], followed by (ii) a *densification* process to ensure that the resulting instance is dense enough to be used for clustering hardness. As the previous best NP-hardness results for clustering [CK19] rely on the hardness of Vertex Coverage [AKS11, Man19, AS19] obtained by recent advances on the 2-to-2 Games Conjecture [KMS17, DKK+18a, DKK+18b, KMS18, BKS19, BK19], we believe that further study on JCH and its variants will lead to more interesting ideas in hardness of approximation that will be useful for understanding geometric optimization problems.

1.1.1 Johnson Coverage Hypothesis

We introduce the Johnson Coverage Hypothesis (JCH) which states that for every $\varepsilon > 0$, there is some $z \in \mathbb{N}$, such that given a parameter $k \in \mathbb{N}$ and a collection E of z-sized subsets (z-sets henceforth) of $[n]$ as input, it is NP-hard to distinguish between the following two cases:

Completeness: There are k subsets of $[n]$ each of cardinality $z-1$, say $S_1, \ldots, S_k$, such that for every $T \in E$ there is some $i \in [k]$ such that $S_i \subset T$.

Soundness: For every k subsets of $[n]$ each of cardinality $z-1$, say $S_1, \ldots, S_k$, we have that there is $E' \subseteq E$ such that

- For every $T \in E'$ and every $i \in [k]$ we have $S_i \not\subset T$.
- $|E'| \geq \left(\frac{1}{e} - \varepsilon\right) \cdot |E|$.

We will often say that a set S *covers* another T if $S \subseteq T$.

We refer to the gap problem described above in JCH as the Johnson Coverage problem. Notice that when $z = 2$, the Johnson Coverage problem is just the Vertex Coverage problem, for which [AS19] have shown that a value close to 0.93 is the correct inapproximability ratio assuming the Unique Games Conjecture. Therefore, for larger values of z, JCH suggests that the Johnson Coverage problem (which is a generalization of the Vertex Coverage problem) gets harder to approximate and approaches the same inapproximability as Max k-Coverage.

Introducing JCH is the main conceptual contribution of this paper. Most previous hardness results for geometric optimization problems [Tre00, GI03, ACKS15, LSW17] use hardness of graph problems in bounded degree graphs and embed them to ℓ_p-metrics, where the

[4]It is interesting to note that approximate nearest-neighbor search, a problem closely related to clustering, is easier to solve in the ℓ_1-metric than general metrics.

bounded degree condition was crucial for the inapproximability ratio. Cohen-Addad and Karthik [CK19], at least in the context of clustering, were the first to show how to embed instances of Vertex Coverage problem (with no restriction on the degree) to obtain hard instances of k-means and k-median.

In this work, we further study embeddability of coverage problems, and show that the embedding of [CK19] can be further generalized to covering problems on hypergraphs, i.e., to the Johnson Coverage problem, which may be as hard as the general covering problem. Johnson Coverage problem is a purely combinatorial covering problem, but the implicit geometric structure of the problem allows us to seamlessly embed it to ℓ_p-metrics.

Plausibility of JCH and Connection to Hypergraph Turán Number. One may reformulate the Johnson Coverage problem as a special case of Max k-Coverage, by asking for instances of Max k-Coverage whose sets and elements correspond to subsets of size $(z-1)$ and z of some universe $[n]$ respectively. So it is natural to study the performance of the standard LP or SDP relaxation, which is closely related to relaxations for the clustering problems.

A natural gap instance for Johnson Coverage problem is the complete z-uniform hypergraph, i.e., $E := \binom{[n]}{z}$. Since each $e \in E$ contains exactly z subsets of size $(z-1)$, the standard LP relaxation and SDP relaxation admit a feasible solution whose value is at most $\binom{[n]}{z-1}/z$ and $\binom{[n]}{z-1}/z-1$ respectively [GL17]. Interestingly, the soundness analysis of this proposed gap instance is closely related to the Hypergraph Turán problem, a long standing open problem in extremal combinatorics [Tur41].

Given $r < z \in \mathbb{N}$, let $\mathsf{Tu}(n,z,r)$ be the minimum $f \in \mathbb{N}$ such that there exists an r-uniform hypergraph $H = ([n], F)$ with n vertices and f edges where every set $T \in \binom{[n]}{z}$ contains at least one hyperedge from F. Let $t(z,r) = \lim_{n\to\infty} \mathsf{Tu}(n,z,r)\binom{n}{r}^{-1}$. The classical Turan's theorem states that $t(z,2) = 1/(z-1)$, with the extremal example being the union of $(z-1)$ equal-sized cliques. However, the quantity of our interest $t(z,z-1)$ is not well understood when $z > 3$; the best lower bound is $1/(z-1)$ [DC83] and the best upper bound is $(1/2+o(1))\ln z/z$ [Sid97], with the conjecture $t(z,z-1) \geq \omega(1/z)$ still open [DC91]. We refer the reader to more recent surveys by Sidorenko [Sid95] and Keevash [Kee11].

What will happen if H has at most a $1/(z-1)$ fraction of hyperedges, which is much less than the conjectured value of $t(z,z-1)$? Then some $T \in \binom{[n]}{z}$ will not be covered by hyperedges of H. If we consider a random hypergraph where each hyperedge is picked with probability $1/z-1$, $S \in \binom{[n]}{z}$ is covered with probability $1-(1-1/z-1)^z$, which converges to $1-1/e$ as z increases. It is thus natural to hypothesize that this is indeed optimal as z increases.

Hypothesis 1.1. *Any $(z-1)$-uniform hypergraph with n vertices and $\binom{n}{z-1}/z-1$ hyperedges covers at most a $d(n,z)$ fraction of sets of size z, where $\lim_{z\to\infty}\lim_{n\to\infty} d(n,z) = 1-1/e$.*

If Hypothesis 1.1 is true, then the afore-proposed gap instance for Johnson Coverage problem (the complete z-uniform hypergraph) has an integrality gap of $(1-1/e+\varepsilon)$ for any $\varepsilon > 0$ for the LP and SDP relaxation. Thus, a refutation of JCH either implies interesting constructions in extremal combinatorics or establishes that there is a polynomial time algorithm outperforming the LP and SDP relaxation, on which the current best approximation algorithms for both k-median and k-means in any metric are based [BPR+15, ANSW20].

At first glance, the hypothesis looks rather strong (i.e., less likely to be true) because of the existence of extremal structures that are strictly better than random hypergraphs. However, these advantages often diminish as the size of the object of interest increases. One example is

the z-clique density in graphs corresponding to $t(z,2)$. After the work of Razborov (for $z = 3$ [Raz08]) and Nikiforov (for $z = 4$ [Nik11]), Reiher [Rei16] proved that the graph that covers the most number of K_z with the given edge density is the union of disjoint cliques of the same size. In our context, since every K_z has $\binom{z}{2}$ edges, the desired edge density is $1/w-1$ with $w = \binom{z}{2}$, so the extremal graph is the union of w disjoint cliques and the probability that a random copy of K_z is not covered by this graph is $\prod_{i=1}^{z}(1 - \frac{i-1}{\binom{z}{2}-1})$, which converges to $1/e$ as z increases. So for large z, the extremal examples do not have an advantage over a random graph! Indeed, this connection already gives improved SDP gaps for various clustering objectives in natural hard instances in ℓ_p-metrics (see Theorem 1.6, or for more details, see Section 3.5).

Technical Barriers. One may wonder if it is possible to start from Feige's hard instances of Max k-Coverage [Fei98] (which have the desired gap of $1 - 1/e$) and simply provide a gap-preserving reduction to Johnson Coverage problem. In Theorem A.1 we showed that if such a reduction exists, then it should significantly blow up the witness size (the number of sets needed to cover the universe in the completeness case). At the heart of this observation is that we are transforming arbitrary set systems (arising from hard instances of Max k-Coverage) to set systems with bounded VC dimension (as in the Johnson Coverage problem). In fact, a reduction in this spirit was recently obtained in [CKL21], and that reduction did indeed blow up the witness size. Therefore the result in [CKL21] where we show Max k-Coverage is hard to approximate beyond $1 - 1/e$ factor on set systems of large girth (thus bounded VC dimension) may be seen as moral progress on understanding JCH.

Additionally, in Theorem A.11, we revisit Feige's framework for showing hardness of approximation of Max k-Coverage, and highlight that certain simple modifications to his reduction would not prove JCH. In particular, we show that no "partition system" can be combined with standard label cover instances to yield JCH. This provides some evidence that in order to prove JCH, we would need to potentially prove hardness of approximation result for some highly structured label cover instances.

Subsequent Work. Motivated by an earlier version of this paper, Guruswami and Sandeep [GS20] initiated the study of the minimization variant of JCH, where the goal is to select as few $(z-1)$-sets as possible to ensure that every z-set in E is covered by a chosen $(z-1)$-set. While the naive algorithm gives a z-approximation, they obtained an $(z/2 + o(z))$-approximation algorithm. They also asked as an open problem whether JCH and hardness of the minimization version can be formally related.

1.1.2 Generalized Johnson Coverage Problem

Motivated by the connections in previous subsubsection to $t(z,r)$ for general $z > r \geq 1$, we consider more generalized Johnson Coverage problems where the input is still a collection of z-sized sets E but instead of choosing $(z-1)$-sized sets, we choose r-sized sets for some $r \in \{1,\ldots,z-1\}$ to cover as many sets in E as possible. (See Definition 3.1 for the formal definition.) For example, when $r = 1$, the problem becomes the z-Hypergraph Vertex Coverage problem. We prove the following for the 3-Hypergraph Vertex Coverage problem:

Theorem 1.2 (Informal statement; See Theorem 4.1)**.** *For every $\varepsilon > 0$, the 3-Hypergraph Vertex Coverage problem is* NP*-Hard to approximate to a factor of $7/8 + \varepsilon$.*

While the covering version of the above problem, namely the *(Minimum) 3-Hypergraph*

Vertex Cover has been studied actively in literature, culminating in [DGKR05], the coverage version does not seem to have been explicitly studied.

We remark that the above inapproximability result is higher than the inapproximability of Vertex Coverage problem proved under UGC (of roughly 0.93) and more so under NP $\neq$ P (of roughly 0.98), and this higher gap will be useful in the next subsubsection for applications.

In Section 3, we provide a embedding framework which converts any hardness result for Generalized Johnson Coverage problem (where $z > r \geq 1$) with hardness ratio $\alpha < 1$ to directly yield inapproximability results for various clustering objectives. Together with Theorem 1.2, this framework produces NP-Hardness results nearly matching or improving the best known hardness results assuming UGC. The case $z = 4$, $r = 2$ also yields improved SDP gaps for various clustering objectives in Section 3.5.

1.1.3 Inapproximability Results for k-means and k-median in ℓ_p-metrics

First, we present our results for 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.

Theorem 1.3 (Discrete version; Informal statement of Theorems 3.12 and 3.13)**.** *Assuming* JCH*, given n points and poly(n) candidate centers, it is* NP*-hard to approximate:*

- *the k-means objective to within a $1 + 8/e \approx 3.94$ factor in ℓ_1-metric and $1 + 2/e \approx 1.73$ factor in ℓ_2-metric.*
- *the k-median objective to within a $1 + 2/e \approx 1.73$ factor in ℓ_1-metric and 1.27 factor in ℓ_2-metric.*

In fact, we obtain inapproximability results for all ℓ_p-metrics (where $p \in \mathbb{R}_{\geq 1}$), and the details are provided in Section 3. Also, the results obtained in the above theorem significantly improves on the bounds of [CK19] (see Table 1). Finally, we note that the bounds obtained for the ℓ_1-metric might be optimal as approximating k-means and k-median to a factor of 3.95 and 1.74 is *fixed parameter tractable* even for general metrics [CGK+19].

We now sketch the proof of Theorem 1.3. We use the notations concerning JCH established in Section 1.1.1. We create a point for each (z size) subset in E and a candidate center for each subset of $[n]$ of size $z - 1$. Both the points and candidate centers are just the characteristic vectors of their corresponding subsets of $[n]$, i.e., the points are all Boolean vectors of Hamming weight z and the candidate centers are all Boolean vectors of Hamming weight $z - 1$. In the completeness case, it is easy to see that there is a set $\mathcal{C}$ of k candidate centers such that every point has a center in $\mathcal{C}$ at (Hamming) distance 1. Also, in the soundness case, it is easy to see that for every set $\mathcal{C}$ of k candidate centers, we have that at least $1/e - \varepsilon$ fraction of the points in E are at (Hamming) distance at least 3 from every center in $\mathcal{C}$. Note that the dimension of this embedding is n, and that we reduce it to $O(\log n)$ by developing a generalization of the dimenionality reduction machinery introduced in [CK19]. Furthermore, the proof from the Hamming metric to other ℓ_p-metrics goes through the composition of various graph embedding gadgets introduced in [CK19] and generalized in this paper to hypergraph embedding.

We now shift our attention to the continuous case, where the centers can be picked at arbitrary locations in the metric space. We show the following[5].

[5]Note that Theorem 1.4 depends on a slight strengthening of JCH, which we call JCH*, where we further assume that for the hard instances of JCH, we have $|E| = \omega(k)$ (see Section 3.4 for further discussion).

Theorem 1.4 (Continuous version; Informal statement of Theorems 3.16, 3.22, 3.24, and 3.25)**.** *Assuming* JCH**, given n points, it is* NP*-hard to approximate:*

- *the* k-means *objective to within* 2.10 *factor in* ℓ_1*-metric and* $1 + 1/e \approx 1.36$ *factor in* ℓ_2*-metric.*
- *the* k-median *objective to within a* $1 + 1/e \approx 1.36$ *factor in* ℓ_1*-metric and* 1.08 *factor in* ℓ_2*-metric.*

The above theorem for k-median in ℓ_1-metric is obtained through an intermediate hardness of approximation proof for k-median in the Hamming metric (see Theorem 3.20). Additionally, the fact that medians and means have nice algebraic definitions in ℓ_1 and ℓ_2 respectively allow us to transfer the hardness from Hamming metrics without much loss. Furthermore, thanks to a technical result of Rubinstein [Rub18] on near isometric embedding from Hamming metric in d dimensions to Edit metric in $O(d \log d)$ dimensions, we can translate all our (discrete case) results in the Hamming metric to the Edit metric (see Appendix A in [CK19] for details).

We also prove the first hardness of approximation results for k-median in ℓ_2-metric (Theorem 3.24) and k-means in ℓ_1-metric (Theorem 3.25). Even though continuous k-median in ℓ_2 has been actively studied for the bounded d or k [ARR98, BHPI02, KSS10, FL11], and to the best of our knowledge, no hardness of approximation for the general case was known in the literature. We remark that independent to our work, in [BGJ21], the authors prove APX-hardness of the Euclidean k-median problem under UGC.

Next, we move our attention to proving NP-Hardness results. With general versions of Theorem 1.3 and Theorem 1.4 for generalized Johnson Coverage problem, Theorem 1.2 proves the following NP-Hardness results presented in Table 1. For continuous versions, more technical work and randomized reductions are needed to ensure enough density (see Section 4.5).

Theorem 1.5. *It is* NP*-hard to approximate the following clustering objectives; discrete* k-means *in* ℓ_1 *within* 1.38*, discrete* k-median *in* ℓ_1 *within* 1.12*, discrete* k-means *in* ℓ_2 *within* 1.17*, discrete* k-median *in* ℓ_2 *within* 1.07*, continuous* k-means *in* ℓ_1 *within* 1.16*, continuous* k-median *in* ℓ_1 *within* 1.06*, continuous* k-means *in* ℓ_2 *within* 1.06*, and continuous* k-median *in* ℓ_2 *within* 1.015*. The results for continuous* k-median *and* k-means *hold under randomized reductions.*

Finally, we present another evidence for usefulness of (generalized) Johnson Coverage problem via the clique density theorem of Reiher [Rei16], which gives improved SDP gaps of $149/125 = 1.192$ for discrete k-median in ℓ_1-metric and k-means in ℓ_2-metric when the instances are *well-separated*. (I.e., points are not closed to one another.) Moreover, this result holds even when the integral solution is allowed to use $\Omega(k)$ more centers. Previously, the best gaps were 1.14 and 1.17 for k-median in ℓ_1-metric and k-means in ℓ_2-metric respectively, both following from the SDP gaps of the Unique Games hardness [CK19].

Theorem 1.6 (Informal version of Theorem 3.26)**.** *Fix any* $\varepsilon > 0$*. For discrete* k-median *in* ℓ_1 *and discrete* k-means *in* ℓ_2*, there is a family of well-separated instances where the SDP relaxation has a gap of at least* $149/125 - \varepsilon \approx 1.192 - \varepsilon$*, even when the integral solution opens* $\Omega(k)$ *more centers.*

1.2 Organization of the Paper

The paper is organized as follows. In Section 2, we introduce some notations that are used throughout the paper, and some tools from coding theory. In Section 3, we introduce JCH, and show how it implies the inapproximability of k-means and k-median in various ℓ_p-metrics (i.e.,

Theorems 1.3 and 1.4). In Section 4 we prove a weak version of JCH. In Section 5 we present some open problems of interest.

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.

Clustering Objectives. Given two sets of points P and C in a metric space, we define the k-means cost of P for C to be $\sum_{p\in P} \left(\min_{c\in C} (\text{dist}(p,c))^2\right)$ and the k-median cost to be $\sum_{p\in P} \left(\min_{c\in C} \text{dist}(p,c)\right)$. Given a set of points P in a metric space and partition π of P into $P_1 \dot\cup P_2 \dot\cup \cdots \dot\cup P_k$, we define the k-minsum cost of P for π to be $\sum_{i\in[k]} \left(\sum_{p,q\in P_i} \text{dist}(p,q)\right)$. Given a set of points P, the k-means/k-median (resp. k-minsum) objective is the minimum over all C (resp. π) of cardinality k of the k-means/k-median (resp. k-minsum) cost of P for C (resp. π). Given a point $p \in P$, the contribution to the k-means (resp. k-median) cost of p is $\min_{c\in C} (\text{dist}(p,c))^2$ (resp. $\min_{c\in C} \text{dist}(p,c)$).

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 denoted by* AG *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.*

An informal argument justifying the existence of the above code family is provided in [CK19]. Furthermore, as noted in [CK19], random codes obtaining weaker parameters than the parameters stated above (see Gilbert-Varshamov bound [Gil52, Var57]) suffice for the results in this paper and it may even be possible to use concatenated codes (arising from Reed-Solomon codes) which approach the Gilbert-Varshamov bound in the proofs in this paper instead of the aforementioned algebraic geometric codes.

3 Conditional Inapproximability of k-means and k-median in ℓ_p-metrics

In this section, we first formally introduce the Johnson Coverage Hypothesis (JCH), then generalize the gadget constructions via graph embedding which were introduced in [CK19], and

finally prove how JCH implies the hardness of approximation results for k-means and k-median, i.e., Theorems 1.3 and 1.4. We also prove improved integrality gaps inspired by JCH (i.e., Theorem 1.6).

3.1 Johnson Coverage Hypothesis

In this subsection, we first introduce the Johnson Coverage problem, followed by the Johnson Coverage hypothesis.

Let $n, z, y \in \mathbb{N}$ such that $n \geq z > y$. Let $E \subseteq \binom{[n]}{z}$ and $S \in \binom{[n]}{y}$. We define the coverage of S w.r.t. E, denoted by $\mathsf{cov}(S, E)$ as follows:

$$\mathsf{cov}(S, E) = \{T \in E \mid S \subset T\}.$$

Definition 3.1 (Johnson Coverage Problem)**.** *In the (α, z, y)-Johnson Coverage problem with $z > y \geq 1$, we are given a universe $U := [n]$, a collection of subsets of U, denoted by $E \subseteq \binom{[n]}{z}$, and a parameter k as input. We would like to distinguish between the following two cases:*

- **Completeness***: There exists $\mathcal{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{y}$ such that*

$$\mathsf{cov}(\mathcal{C}) := \bigcup_{i \in [k]} \mathsf{cov}(S_i, E) = E.$$

- **Soundness***: For every $\mathcal{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{y}$ we have $|\mathsf{cov}(\mathcal{C})| \leq \alpha \cdot |E|$.*

We call $(\alpha, z, z-1)$-Johnson Coverage as (α, z)-Johnson Coverage.

Notice that $(\alpha, 2)$-Johnson Coverage Problem is simply the well-studied vertex coverage problem (with gap α). Also, notice that if instead of picking the collection $\mathcal{C}$ from $\binom{[n]}{y}$, we replace it with picking the collection $\mathcal{C}$ from $\binom{[n]}{1}$ with a similar notion of coverage, then we simply obtain the Hypergraph Vertex Coverage problem (which is equivalent to the Max k-Coverage problem for unbounded z).

We now put forward the following hypothesis.

Hypothesis 3.2 (Johnson Coverage Hypothesis (JCH))**.** *For every constant $\varepsilon > 0$, there exists a constant $z := z(\varepsilon) \in \mathbb{N}$ such that deciding the $\left(1 - \frac{1}{e} + \varepsilon, z\right)$-Johnson Coverage Problem is* NP*-Hard.*

Since Vertex Coverage problem is a special case of the Johnson Coverage problem, we have that the NP-Hardness of (α, z)-Johnson Coverage problem is already known for $\alpha = 0.94$ [AS19] (under unique games conjecture).

On the other hand, if we replace picking the collection $\mathcal{C}$ from $\binom{[n]}{z-1}$ by picking from $\binom{[n]}{1}$, then for the Hypergraph Vertex Coverage problem, we do know that for every $\varepsilon > 0$ there is some constant z such that the Hypergraph Vertex Coverage problem is NP-Hard to decide for a factor of $\left(1 - \frac{1}{e} + \varepsilon\right)$ [Fei98].

Related work to JCH**.** Johnson Coverage problem can be considered as a special case of Max k-Coverage in set systems with a natural additional structure. Such a restriction of fundamental covering / packing / constraint satisfaction problems to structured instances often arise in geometric settings such as Independent Set of Rectangles [CE16]. The *VC dimension* has

been an important combinatorial notion capturing many implicit structures posed by geometric problems that allow better algorithms [BG95, BKL12] than general set systems in some regimes. Recently, Alev et al. [AJT19] studied approximating constraint satisfaction problems on a high-dimensional expander, another object that bridges graph theory and geometry, and showed that this structure makes CSPs easier to approximate.

3.2 Gadget Constructions via Graph Embeddings

In this subsection, we recall a notion of graph embedding that was introduced in [CK19]. And then we prove some bounds on the embedding for important ℓ_p-metrics. These embedding are then used in the next subsection to prove inapproximability results.

Let $q, t, r \in \mathbb{N}$ such that $q \geq t \geq r$. Let $J(q,t,r)$ denote the incidence graph of the Johnson graph [HS93]. Elaborating, we define $J(q,t,r)$ to be the bipartite graph on partite sets $\binom{[q]}{t}$ and $\binom{[q]}{r}$ where we have an edge $(S, S') \in \binom{[q]}{t} \times \binom{[q]}{r}$ in $J(q,t,r)$ if and only if $S' \subseteq S$. We use the shorthand $J(q,t)$ to denote $J(q,t,t-1)$.

We would like to analyze the embedding of $J(q,t)$ into ℓ_p-metric spaces for all $p \in \mathbb{R}_{\geq 1}$.

Definition 3.3 (Gap Realization of a Bipartite graph [CK19])**.** *Let $p \in \mathbb{R}_{\geq 1}$. 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 [KM20] and is closely related to notion of contact dimension which has been well-studied in literature since the early eighties [Pac80, Mae85, FM88, Mae91, DKL19]. We refer the reader to [CK19] for further discussion.

Definition 3.4 (Gap number)**.** *Let $p \in \mathbb{R}_{\geq 1}$. 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[6] where $d \leq poly(|A| + |B|)$.*

In [CK19], the authors studied $g_p(J(q,t,1))$. In this paper, we are interested in analyzing $g_p(J(q,t,t-1))$ for all $q, t \in \mathbb{N}$ ($q \geq t$) and $p \in \mathbb{R}_{\geq 1}$. We remark that we do not study $g_\infty(J(q,t,t-1))$ in this paper, as this was already settled to be equal to 3 in [CK19]. We recall the following upper bound on gap number which follows immediately from triangle inequality:

Proposition 3.5 (Essentially [CK19])**.** *Let $q \geq 3$, $t \geq 2$ (where $q \geq t$), and $p \in \mathbb{R}_{\geq 1}$. We have $g_p(J(q,t)) \leq 3$.*

We consider the ℓ_1-metric and show that we can meet the upper bound in Proposition 3.5.

[6] 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.

Lemma 3.6. *For all $q \geq 3$ and $t \geq 2$ (where $q \geq t$), we have $g_1(J(q,t)) = 3$. More generally, $g_1(J(q,t,s)) = (t-s+2)/(t-s)$.*

Proof. For the ℓ_1-metric consider the mapping $\tau : \binom{[q]}{t} \cup \binom{[q]}{s} \to \{0,1\}^q$ defined as follows. For every $S \in \binom{[q]}{t} \cup \binom{[q]}{s}$, we define

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

Fix some $(S, S') \in \binom{[q]}{t} \times \binom{[q]}{s}$ such that $S' \subseteq S$. Then we have that

$$\tau(S) - \tau(S') = \sum_{i \in S \setminus S'} e_i \Rightarrow \|\tau(S) - \tau(S')\|_1 = t - s.$$

On the other hand if we fix some $(S, S') \in \binom{[q]}{t} \times \binom{[q]}{s}$ such that $S' \not\subset S$ then we have that

$$\|\tau(S) - \tau(S')\|_1 = \left\| \left(\sum_{i \in S \setminus S'} e_i \right) - \left(\sum_{i \in S' \setminus S} e_i \right) \right\|_1 \geq t - s + 2.$$

Thus we have that τ, $(t-s+2)/(t-s)$-gap-realizes $J(q,t,s)$ in the ℓ_1-metric. □

Now we focus our attention to bounding the gap number in the Euclidean metric. First, we see that the below lower bound simply follows from τ constructed in the above proof.

Corollary 3.7. *For all $q \geq 3$ and $t \geq 2$ (where $q \geq t$), we have $g_2(J(q,t)) \geq \sqrt{3}$ and $g_2(J(q,t)) \geq \sqrt{(t-s+2)/(t-s)}$.*

However, we improve the lower bound with a different embedding.

Lemma 3.8. *For all $q \geq 3$ and $t \geq 2$ (where $q \geq t$), we have $g_2(J(q,t,1)) \geq \sqrt{1 + \frac{1}{\sqrt{t}-1}}$. More generally, $g_2(J(q,t,s)) \geq \sqrt{1 + \frac{1}{\sqrt{ts}-s}}$.*

Proof. For the ℓ_2-metric consider the mapping $\tau : \binom{[q]}{t} \cup \binom{[q]}{s} \to \mathbb{R}^q$ defined as follows. For every $T \in \binom{[q]}{t}$, we define

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

For every $S \in \binom{[q]}{s}$, we define

$$\tau(S) = \sqrt{\frac{t}{s}} \cdot \sum_{i \in S} e_i.$$

Fix some $(T, S) \in \binom{[q]}{t} \times \binom{[q]}{s}$ such that $S \subseteq T$. Then we have that

$$\tau(T) - \tau(S) = \left(\left(\sqrt{\frac{t}{s}} - 1 \right) \cdot \sum_{i \in S} e_i \right) + \left(\sum_{i \in T \setminus S} e_i \right)$$

$$\Rightarrow \|\tau(T) - \tau(S)\|_2 = \sqrt{s \cdot \left(\sqrt{\frac{t}{s}} - 1 \right)^2 + t - s}$$

$$= \sqrt{2} \cdot \sqrt{t - \sqrt{ts}}$$

On the other hand if we fix some $(T, S) \in \binom{[q]}{t} \times \binom{[q]}{s}$ such that $S \not\subset T$ then we have that

$$\begin{aligned}
\|\tau(T) - \tau(S)\|_2 &= \left\| \left(\left(\sqrt{\frac{t}{s}} - 1 \right) \cdot \sum_{i \in S \cap T} e_i \right) + \left(\sqrt{\frac{t}{s}} \cdot \sum_{i \in S \setminus T} e_i \right) + \left(\sum_{i \in T \setminus S} e_i \right) \right\|_2 \\
&\geq \sqrt{(s-1) \cdot \left(\sqrt{\frac{t}{s}} - 1 \right)^2 + t - s + 1 + \frac{t}{s}} \\
&= \sqrt{2} \cdot \sqrt{t - \sqrt{ts} + \sqrt{\frac{t}{s}}} \\
&= \sqrt{2} \cdot \sqrt{(t - \sqrt{ts}) \cdot \left(1 + \frac{1}{\sqrt{ts} - s} \right)}
\end{aligned}$$

Thus we have that τ, $\sqrt{\left(1 + \frac{1}{\sqrt{ts}-s}\right)}$-gap-realizes $J(q,t,s)$ in the ℓ_2-metric. When $s = 1$, τ $\sqrt{\left(1 + \frac{1}{\sqrt{t}-1}\right)}$-gap-realizes $J(q,t,1)$ in the ℓ_2-metric. □

In order to see that the lower bound on $g_2(J(q,t,s))$ given in Lemma 3.8 is indeed higher than the one given in Corollary 3.7, note the following:

$$\left(1 + \frac{1}{\sqrt{ts} - s}\right) - \left(\frac{t - s + 2}{t - s}\right) = \frac{1}{\sqrt{ts} - s} - \frac{2}{t - s} = \frac{t - s - 2\sqrt{ts} + 2s}{(\sqrt{ts} - s)(t - s)} = \frac{(\sqrt{t} - \sqrt{s})^2}{(\sqrt{ts} - s)(t - s)} > 0,$$

as $t > s$.

We wrap up our computation of gap numbers by showing that as p grows the gap number of $J(q,t)$ in the ℓ_p-metric approaches 3.

Lemma 3.9. *For all $q \geq 3$ and $t \geq 2$ (where $q \geq t$), we have that for every $\varepsilon > 0$ there exists $p \in \mathbb{N}$ such that $g_p(J(q,t)) > 3 - \varepsilon$.*

Proof. Fix $q \geq 3$, $t \geq 2$, and $\varepsilon > 0$. Let $p \in \mathbb{N}$ such that $q^{1/p} < 1 + \varepsilon/3$. Consider the mapping $\tau : \binom{[q]}{t} \cup \binom{[q]}{t-1} \to \mathbb{R}^q$ defined as follows. For every $S \in \binom{[q]}{t-1}$, we define

$$\tau(S) = \left(\frac{\vec{1}}{2}\right) + \sum_{i \in S} e_i,$$

and for every $T \in \binom{[q]}{t}$, we define

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

Fix some $(S, T) \in \binom{[q]}{t-1} \times \binom{[q]}{t}$ such that $S \subset T$. Then we have that

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

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

On the other hand if we fix some $(S, T) \in \binom{[q]}{t-1} \times \binom{[q]}{t}$ such that $S \not\subset T$, i.e., $\exists u \in [q]$ such that $u \in S \setminus T$ then we have that

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

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

We remark that the embeddings given in Lemmas 3.6 and 3.9 are essentially small modifications of the ones given in [CK19].

For the sake of compactness of statements in the future, we introduce the following.

Definition 3.10. *For all $p \in \mathbb{R}_{\geq 1}$, we define $\gamma_p = \inf_{q \geq t} g_p(J(q,t))$, and $\gamma_{p,\Delta} = \inf_{q \geq t \geq \Delta+1} g_p(J(q,t,t-\Delta))$*

Finally, we conclude this subsection by recalling the following 'hereditary' property of the embedding which is important for applications to hardness of approximation.

Proposition 3.11 ([CK19])**.** *Let $q \geq 3$ and $t \geq 2$ (where $q \geq t$), and $p \in \mathbb{R}_{\geq 1}$. Let $E \subseteq \binom{[q]}{t}$ and G be the subgraph of $J(q,t)$ induced by the vertex set $E \cup \binom{[q]}{t-1}$. Let τ be a λ-gap realization of $J(q,t)$ in the ℓ_p-metric. Then τ restricted to the vertices of G is a λ-gap realization of G in the ℓ_p-metric.*

3.3 Conditional Inapproximability of Discrete k-means and k-median

In this subsection, we show how JCH or any hardness of (α, z, y)-Johnson Coverage implies the hardness of approximation of the discrete case of k-means and k-median in all ℓ_p-metrics, i.e., we prove Theorem 1.3.

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

$$\zeta_1(p, \Delta, \alpha) := 1 + (1-\alpha)(\gamma_{p,\Delta} - 1) \text{ and } \zeta_2(p, \Delta, \alpha) := 1 + (1-\alpha)(\gamma_{p,\Delta}^2 - 1).$$

Also let $\zeta_1(p) := \zeta_1(p, 1, 1 - 1/e)$ and $\zeta_2(p) := \zeta_2(p, 1, 1 - 1/e)$. Notice that $\zeta_1(1) \approx 1.73$, $\zeta_2(1) = 3.94$, $\zeta_1(2) \approx 1.27$, $\zeta_2(2) \approx 1.73$, and as $p \to \infty$, we have $\zeta_1(p) \to \zeta_1(1)$ and $\zeta_2(p) \to \zeta_2(1)$.

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

Theorem 3.12 (k-means with candidate centers in $O(\log n)$ dimensional ℓ_p-metric space)**.** *Let $p \in \mathbb{R}_{\geq 1}$. Assuming (α, z, y)-Johnson Coverage is NP-hard, for every constant $\varepsilon > 0$, 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*

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

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

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

for some constant $\rho > 0$.

In particular, assuming JCH, k-means is NP-Hard to approximate within a factor $\zeta_2(p)$, which is at least 3.94 in ℓ_1 and at least 1.73 in ℓ_2.

Theorem 3.13 (k-median with candidate centers in $O(\log n)$ dimensional ℓ_p-metric space)**.** *Let* $p \in \mathbb{R}_{\geq 1}$. *Assuming* (α, z, y)*-Johnson Coverage is* NP*-Hard, for every constant* $\varepsilon > 0$, *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*

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

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

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

for some constant $\rho > 0$.

In particular, assuming JCH, k-median is NP-Hard to approximate within a factor $\zeta_2(p)$, which is at least 1.73 in ℓ_1 and at least 1.27 in ℓ_2.

Remark 3.14. *Notice that for every* q, z, *and* p, *from the embedding given by Lemma 3.6, we can deduce that* $g_p(J(q, z, 1))$ *is a constant (depending on* q, z, *and* p*) bounded away from 1. Thus we have for* $i \in \{1, 2\}$ *that* $\zeta_i(p, z-1, \alpha) \geq 1 + \delta$ *for some positive constant* δ *depending only on* p, α, *and* z. *On the other hand, Feige [Fei98] showed that for any constant* $\varepsilon > 0$ *there is a constant* $z := z(\varepsilon)$ *such that* $(1 - 1/e + \varepsilon, z, 1)$*-Johnson Coverage problem is* NP*-Hard. Plugging these parameters in Theorems 3.12 and 3.13, we have that* k-means *and* k-median *are* APX*-Hard in all* ℓ_p*-metrics. In fact, we do not even need to invoke the tight result of [Fei98], as it's relatively easy to show that* $(1 - \delta, z, 1)$*-Johnson Coverage problem is* NP*-Hard directly from the PCP theorem [AS98, ALM+98, Din07] for some constants* $\delta > 0$ *and* $z \in \mathbb{N}$.

The proof of the above two theorems follows by generalizing the embedding framework established in [CK19]. In [CK19], the authors considered a two-player communication game between Alice and Bob, where Alice holds as input an edge in a publicly known graph and Bob holds as input a vertex in the same graph, and the goal is for Bob to send a message to Alice (using public randomness), so that Alice can decide if her edge is incident on Bob's vertex. The authors then showed how an efficient randomized protocol for this communication problem can be combined with a certain embedding of the complete graph (into ℓ_p-metrics) to obtain inapproximaility results for k-means and k-median in ℓ_p-metrics.

Below we consider the game where Alice holds a z-sized subset of $[n]$ and Bob holds a y-sized subset of $[n]$, and the goal is for Bob to send a message to Alice (using public randomness), so that Alice can decide if her subset completely contains Bob's subset. We then design

an efficient randomized protocol for this communication problem using Algebraic Geometric codes (as in [CK19]) and show how the transcript of the protocol can be combined with a certain embedding of the Johnson graph (into ℓ_p-metrics) to obtain the two theorems. We skip providing further details about the framework for the sake of brevity, and point the reader to [CK19].

Proof of Theorems 3.12 and 3.13. Fix $\varepsilon > 0$ as in the theorem statement. Also, fix $p \in \mathbb{R}_{\geq 1}$. Let $\varepsilon' := \varepsilon/11$ (setting ε' to be $\varepsilon/4$ suffices for Theorem 3.13). Starting from a hard instance of (α, z, y)-Johnson Coverage problem (U, E, k), we create an instance of the k-means, or of the k-median problem using Algebraic-Geometric codes and the embedding given in Proposition 3.11 as follows. Let q be a (constant) prime square greater than $\left(\frac{18z^2}{\varepsilon'}\right)^2$. Let AG be the code guaranteed by Theorem 2.1 over alphabet of size q of message length $\eta := \log_q n$ (recall $|U| = n$), block length $\ell := O_q(\eta)$, and relative distance at least $1 - 3/\sqrt{q}$. Let τ be a $g_p(J(q,z,y))$-gap realization embedding of $J(q,z,y)$ which maps vertices of $J(q,z,y)$ to a d^*-dimensional space (note that q, z, and y are all constants, and thus so is d^*). Let $\beta > 0$ be the constant from Definition 3.3.

Construction. The k-median or k-means instance consists of the set of candidate centers $\mathcal{C} \subseteq \mathbb{R}^{\ell \cdot d^*}$ and the set of points to be clustered $\mathcal{P} \subseteq \mathbb{R}^{\ell \cdot d^*}$ which will be defined below. First, we define functions $A_E : E \times [\ell] \to \binom{[q]}{z} \cup \{\perp\}$ and $A_F : \binom{[n]}{y} \times [\ell] \to \binom{[q]}{y} \cup \{\perp\}$ below. Then, we will construct functions $\widetilde{A}_E : E \to \mathbb{R}^{d^* \cdot \ell}$ and $\widetilde{A}_F : \binom{[n]}{y} \to \mathbb{R}^{d^* \cdot \ell}$. Given $\widetilde{A}_E$ and $\widetilde{A}_F$ the point-set $\mathcal{P}$ is just defined to be

$$\left\{ \widetilde{A}_E(T) \middle| T \in E \right\},$$

and the set of candidate centers $\mathcal{C}$ is just defined to be

$$\left\{ \widetilde{A}_F(S) \middle| S \in \binom{[n]}{y} \right\}.$$

For every $\gamma \in [\ell]$ and every $S \in \binom{[n]}{y}$, we define $R^y_{S,\gamma} \subseteq [q]$, where $\mu \in [q]$ is contained in $R^y_{S,\gamma}$ if and only if there exists some $u \in S$ such that[7] $\mathsf{AG}(u)_\gamma = \mu$. Then, we define $A_F(S,\gamma) = R^y_{S,\gamma}$ if $|R^y_{S,\gamma}| = y$ and $A_F(S,\gamma) = \perp$ otherwise. Similarly, for every $\gamma \in [\ell]$ and every $T \in E \subseteq \binom{[n]}{z}$ we define $R^z_{T,\gamma} \subseteq [q]$, where $\mu \in [q]$ is contained in $R_{T,\gamma}$ if and only if there exists some $u \in T$ such that $\mathsf{AG}(u)_\gamma = \mu$. Then, we define $A_E(T,\gamma) = R^z_{T,\gamma}$ if $|R^z_{T,\gamma}| = z$ and $A_E(T,\gamma) = \perp$ otherwise.

For every $x, y \in [q]$ such that $x < x'$ we define $\Lambda_{x,x'} : \binom{[q]}{x} \to \binom{[q]}{x'}$ as follows: For every $X \in \binom{[q]}{x}$ define Δ_X to be the set of $(x' - x)$ many smallest integers in $[q]$ not contained in X. Then $\Lambda_{x,x'}(X) = X \cup \Delta_X$. Now we can construct functions $\widetilde{A}_E : E \to \mathbb{R}^{d^* \cdot \ell}$ and $\widetilde{A}_F : \binom{[n]}{x'} \to \mathbb{R}^{d^* \cdot \ell}$ as follows:

$$\forall \gamma \in [\ell],\ \widetilde{A}_E(T)|_\gamma = \begin{cases} \tau(A_E(T,\gamma)) \text{ if } A_E(T,\gamma) \neq \perp \\ \tau(\Lambda_{|R^z_{T,\gamma}|,z}(R^z_{T,\gamma})) \text{ otherwise} \end{cases}$$

$$\text{and} \quad \widetilde{A}_F(S)|_\gamma = \begin{cases} \tau(A_F(S,\gamma)) \text{ if } A_F(S,\gamma) \neq \perp \\ \tau(\Lambda_{|R^{x'}_{S,\gamma}|,x'}(R^{x'}_{S,\gamma})) \text{ otherwise} \end{cases}.$$

[7] Here we think of γ as a field element of $\mathbb{F}_q$ by using some canonical bijection between $[q]$ and $\mathbb{F}_q$.

Structural Observations. Fix $S \in \binom{[n]}{y}$. We define the set L_S^y as follows:

$$L_S^y := \{\gamma \in [\ell] : |R_{S,\gamma}^y| = y\}.$$

Consider the set $W_S = \{\mathsf{AG}(u) \mid u \in S\}$. Since any two codewords of AG agree on at most $3/\sqrt{q}$ fraction of coordinates, we have by union bound that there are at least $1 - \binom{y}{2}(3/\sqrt{q})$ fraction of coordinates of $[\ell]$ on which all codewords in W_S are distinct. Therefore, we have that $|L_S^y| \geq (1 - \binom{y}{2}(3/\sqrt{q})) \cdot \ell$. Also, note that for all $\gamma \in L_S^y$ we have $A_F(S, \gamma) \neq \perp$.

Similarly, we fix $T \in E$. We define the set L_T^z as follows:

$$L_T^z := \{\gamma \in [\ell] : |R_{T,\gamma}^z| = z\}.$$

By following the averaging argument above, we also have that $|L_T^z| \geq (1 - \binom{z}{2}(3/\sqrt{q})) \cdot \ell$. Again, note that for all $\gamma \in L_T^z$ we have $A_E(T, \gamma) \neq \perp$.

Finally, for every $(S, T) \in \binom{[n]}{y} \times E$, we define $L_{S,T} := L_S^y \cap L_T^z$, and note the following

$$|L_{S,T}| \geq \left(1 - \frac{3zy}{\sqrt{q}}\right) \cdot \ell.$$

Let us now compute a few distances. Consider $(S, T) \in \binom{[n]}{y} \times E$ such that $S \subset T$ then we have

$$\|\widetilde{A}_E(T) - \widetilde{A}_F(S)\|_p = \beta \cdot \ell^{1/p}. \tag{1}$$

This is because, if $S \subset T$ then for all $\gamma \in [\ell]$ we have $R_{S,\gamma}^y \subseteq R_{T,\gamma}^z$. Fix $\gamma \in [\ell]$. If $\gamma \in L_{S,T}$ then we have that $\|\widetilde{A}_E(T)|_\gamma - \widetilde{A}_F(S)|_\gamma\|_p^p = \beta^p$ as $(R_{S,\gamma}^y, R_{T,\gamma}^z)$ is an edge in $J(q, z)$. On the other hand if $\gamma \notin L_T^z$ then either $\gamma \notin L_S^y$ and we have that $\Lambda_{|R_{S,\gamma}^y|,y}(R_{S,\gamma}^y) \subseteq \Lambda_{|R_{T,\gamma}^z|,z}(R_{T,\gamma}^z)$ because $R_{S,\gamma}^y \subseteq R_{T,\gamma}^z$, or we have $\gamma \in L_S^y$, in which case, we have $R_{T,\gamma}^z = R_{S,\gamma}^y$ and thus $R_{S,\gamma}^y \subseteq \Lambda_{|R_{T,\gamma}^z|,z}(R_{T,\gamma}^z)$. In the former case, we have that $(\Lambda_{|R_{S,\gamma}^y|,y}(R_{S,\gamma}^y), \Lambda_{|R_{T,\gamma}^z|,z}(R_{T,\gamma}^z))$ is an edge in $J(q, z)$ and in the latter case we have that $(R_{S,\gamma}^y, \Lambda_{|R_{T,\gamma}^z|,z}(R_{T,\gamma}^z))$ is an edge in $J(q, z)$. The final result in both cases is that $\|\widetilde{A}_E(T)|_\gamma - \widetilde{A}_F(S)|_\gamma\|_p^p = \beta^p$. Therefore, we have

$$\sum_{\gamma \in [\ell]} \|\widetilde{A}_E(T)|_\gamma - \widetilde{A}_F(S)|_\gamma\|_p^p = \|\widetilde{A}_E(T) - \widetilde{A}_F(S)\|_p^p = \beta^p \cdot \ell.$$

Consider $(S, T) \in \binom{[n]}{y} \times E$ such that $S \not\subset T$. Let $u \in U$ such that $u \in S$ but $u \notin T$. We define $\mathsf{Good} := \{\gamma \in [\ell] : \mathsf{AG}(u)_\gamma \notin R_{T,\gamma}^z\}$. Since the relative distance of AG is $1 - 3/\sqrt{q}$, we have by simple union bound that

$$|\mathsf{Good}| \geq \left(1 - \frac{3z}{\sqrt{q}}\right) \cdot \ell.$$

Then, we have

$$\|\widetilde{A}_E(T) - \widetilde{A}_F(S)\|_p \geq (g_p(J(q, y, z)) - \delta) \cdot \beta \cdot \ell^{1/p}, \tag{2}$$

where $\delta := \frac{18zy}{\sqrt{q}}$ (note that $\delta \leq \varepsilon'$). To see this first observe that:

$$\|\widetilde{A}_E(T) - \widetilde{A}_F(S)\|_p^p \geq \sum_{\gamma \in L_{S,T} \cap \mathsf{Good}} \|\widetilde{A}_E(T)|_\gamma - \widetilde{A}_F(S)|_\gamma\|_p^p.$$

Fix $\gamma \in L_{S,T} \cap \mathsf{Good}$. Notice that $R^y_{S,\gamma} \not\subset R^z_{T,\gamma}$ and thus $(R^y_{S,\gamma}, R^z_{T,\gamma})$ is not an edge in $J(q,z)$. Therefore, we have

$$\|\widetilde{A}_E(T)|_\gamma - \widetilde{A}_F(S)|_\gamma\|_p^p \geq g_p(J(q,z,y))^p \cdot \beta^p.$$

Since $|L_{S,T} \cap \mathsf{Good}| \geq (1 - 3z/\sqrt{q} - 3zy/\sqrt{q}) \cdot \ell \geq (1 - 6zy/\sqrt{q}) \cdot \ell$, (2) follows immediately by noting that (i) $g_p(J(q,z,y)) \leq 3$, and (ii) for any $\eta \in [0,1]$, $(1-\eta)^{1/p} \geq (1-\eta)$.

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

Completeness. Suppose there exist $S_1, \ldots, S_k \in \binom{[n]}{y}$ such that $\bigcup_{i \in [k]} \mathsf{cov}(S_i, E) = E$. Then, we define $\mathcal{C}' = \{\widetilde{A}_F(S_i) \mid i \in [k]\}$ and we define $\sigma : \mathcal{P} \to \mathcal{C}'$ as follows: for every $a \in \mathcal{P}$, where $a := \widetilde{A}_E(T)$ for some $T \in E$, let $\sigma(a)$ be equal to $\widetilde{A}_F(S_i)$ such that $S_i \subset T$ (if there is more than one $i \in [k]$ for which S_i is contained in T then we choose one arbitrarily). Therefore for any $a \in \mathcal{P}$ we have from (1)

$$\|a - \sigma(a)\|_p = \beta \cdot \ell^{1/p}.$$

The k-means cost of the overall instance is thus $\beta^2 \cdot \ell^{2/p} \cdot |\mathcal{P}|$, while the k-median cost is $\beta \cdot \ell^{1/p} \cdot |\mathcal{P}|$. 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 (and that $\sigma : \mathcal{P} \to \mathcal{C}'$ simply maps to the closest center in $\mathcal{C}'$). Let $\mathcal{S} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{y}$ be the collection of y sized subsets of U corresponding to the centers of $\mathcal{C}'$, namely

$$\mathcal{S} = \left\{ S_i \in \binom{[n]}{y} \Big| \widetilde{A}_F(S_i) \in \mathcal{C}' \right\}.$$

By the assumptions of the soundness case, we have $\mathsf{cov}(\mathcal{S}) \leq \alpha \cdot |E|$. For each $T \in \mathsf{cov}(\mathcal{S})$, from (1), we have that the contribution of $\widetilde{A}_E(T)$ to the k-means cost is exactly $\beta^2 \cdot \ell^{2/p}$, and to the k-median cost is exactly $\beta \cdot \ell^{1/p}$. However, from (2), for any other $T \in E \setminus \mathsf{cov}(\mathcal{S})$, the contribution of $\widetilde{A}_E(T)$ to the k-median and k-means cost is (at least) respectively $(g_p(J(q,y,z)) - \delta) \cdot \beta \cdot \ell^{1/p}$ and $(g_p(J(q,y,z)) - \delta)^2 \cdot \beta^2 \cdot \ell^{2/p}$. Therefore, the optimal solution w.r.t. k-median objective has cost at least:

$$\begin{aligned}
&\alpha \cdot |E| \cdot \beta \cdot \ell^{1/p} + (1-\alpha) \cdot |E| \cdot (\gamma_{p,z-y} - \delta) \cdot \beta \cdot \ell^{1/p} \\
&\geq |\mathcal{P}| \cdot \beta \cdot \ell^{1/p} \cdot (\zeta_1(p, z-y, \alpha) - \delta) \\
&\geq |\mathcal{P}| \cdot \beta \cdot \ell^{1/p} \cdot (\zeta_1(p, z-y, \alpha) - \varepsilon).
\end{aligned}$$

Similarly, the optimal solution w.r.t. k-means objective has cost at least:

$$\begin{aligned}
&\alpha \cdot |E| \cdot \beta^2 \cdot \ell^{2/p} + (1-\alpha) \cdot |E| \cdot (\gamma_{p,z-y} - \delta)^2 \cdot \beta^2 \cdot \ell^{2/p} \\
&\geq |\mathcal{P}| \cdot \beta^2 \cdot \ell^{2/p} \cdot (\zeta_2(p, z-y, \alpha) - 2\delta) \\
&\geq |\mathcal{P}| \cdot \beta^2 \cdot \ell^{2/p} \cdot (\zeta_2(p, z-y, \alpha) - \varepsilon).
\end{aligned}$$

□

We remark that the above proof can be made significantly notation-light, if we wanted to prove Theorem 1.3 in high dimensions (i.e., $\text{poly}(n)$ dimensions). In particular, we could skip the use of Algebraic Geometric codes. However, the result is more interesting when the dimension is $O(\log n)$, as proving the same NP-Hardness for $(\log n)^{1-o(1)}$ dimensions would violate the Exponential Time Hypothesis [IP01, IPZ01], due to the sub-exponential time approximation algorithm of [Coh18] for such dimensions.

3.4 Conditional Inapproximability of Continuous k-means and k-median

In this subsection, we show how JCH or any hardness of (α, z, y)-Johnson Coverage implies the hardness of approximation of various clustering objectives in the continuous case and i.e., we prove Theorem 1.4. In order to prove the results in this section, we need to assume a slight strengthening of JCH.

Hypothesis 3.15 (Dense Johnson Coverage Hypothesis (JCH*))**.** JCH *holds for instances* (U, E, k) *of Johnson Coverage problem where* $|E| = \omega(k)$.

More generally, let (α, z, y)-Johnson Coverage* be the special case (α, z, y)-Johnson Coverage where the instances satisfy $|E| = \omega(k \cdot |U|^{z-y-1})$. Then JCH* states that for any $\varepsilon > 0$, there exists $z = z(\varepsilon)$ such that $(1 - 1/e + \varepsilon, z, z - 1)$-Johnson Coverage* is NP-Hard. This additional property has always been obtained in literature by looking at the hard instances that were constructed. In [CK19], where the authors proved the previous best inapproximability results for continuous case k-means and k-median, it was observed that hard instances of $(0.94, 2, 1)$-Johnson Coverage constructed in [AKS11, AS19] can be made to satisfy the above property.

First, we consider Euclidean k-means.

Theorem 3.16 (k-means without candidate centers in $O(\log n)$ dimensional Euclidean space)**.** *Assume* (α, z, y)*-Johnson Coverage** *is* NP*-Hard. For every constant* $\varepsilon > 0$*, 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*

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

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

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_2^2 \geq \left(1 + \frac{(1-\alpha)}{(z-y)} - \varepsilon\right) \cdot \rho n \log n,$$

for some constant $\rho > 0$.

Proof. The construction of the point-set $\mathcal{P}$ in the theorem statement is the same as in the proof of Theorem 3.12 (with a further simplification as we fix the embedding given in Lemma 3.8). However, the soundness analysis to prove the theorem is more intricate.

Fix $\varepsilon > 0$ as in the theorem statement. Also, fix $p \in \mathbb{R}_{\geq 1}$. Let $\varepsilon' := \varepsilon/11$. Starting from a hard instance of (α, z, y)-Johnson Coverage* problem (U, E, k) with $|U| = n$ and $|E| = \omega(n^{z-y})$,

we create an instance of the k-means using Algebraic-Geometric codes and the embedding given in Lemma 3.8 as follows. Let q be a (constant) prime square greater than $\left(\frac{18z^2}{\varepsilon'}\right)^2$. Let AG be the code guaranteed by Theorem 2.1 over alphabet of size q of message length $\eta := \log_q n$ (recall $|U| = n$), block length $\ell := O_q(\eta)$, and relative distance at least $1 - 3/\sqrt{q}$.

Construction. The k-means instance consists of the set of points to be clustered $\mathcal{P} \subseteq \{0,1\}^{\ell \cdot q}$ which will be defined below. Recall the function $A_E : E \times [\ell] \to \binom{[q]}{z} \cup \{\perp\}$ defined in the proof of Theorem 3.12 (also recall the definitions $R^z_{T,\gamma}$, $\Lambda_{x,y}$, L^z_T, $\widetilde{A}_F$). We construct $\widetilde{A}_E : E \to \{0,1\}^{q \cdot \ell}$ below. Given $\widetilde{A}_E$ the point-set $\mathcal{P}$ is just defined to be

$$\left\{\widetilde{A}_E(T) \middle| T \in E\right\}.$$

Let τ map a subset of $[q]$ to its characteristic vector in $\{0,1\}^q$. Now we can construct $\widetilde{A}_E$ as follows:

$$\forall \gamma \in [\ell],\ \widetilde{A}_E(T)|_\gamma = \begin{cases} \tau(A_E(T,\gamma)) \text{ if } A_E(T,\gamma) \neq \perp \\ \tau(\Lambda_{|R^z_{T,\gamma}|,z}(R^z_{T,\gamma})) \text{ otherwise} \end{cases}.$$

Structural Observations. We compute distances in a couple of cases. Consider distinct $T, T' \in E$ such that $|T \cap T'| = y$. Then we have

$$(2(z-y) - \delta) \cdot \ell \leq \|\widetilde{A}_E(T) - \widetilde{A}_E(T')\|_2^2 \leq 2(z-y)\ell, \tag{3}$$

where $\delta := 18z^2/\sqrt{q}$ (note that $\delta \leq \varepsilon'$).

Now consider distinct $T, T' \in E$ such that $|T \cap T'| < y$. Then we have

$$(2(z-y) + 2 - \delta) \cdot \ell \leq \|\widetilde{A}_E(T) - \widetilde{A}_E(T')\|_2^2. \tag{4}$$

We now analyze the k-means cost of the instance. To do so, we will make use of the following classic fact about the k-means cost of a partition $C_1 \dot\cup \cdots \dot\cup C_k = \mathcal{P}$ of a set of points in $\mathbb{R}^d$.

Fact 3.17. *Given a clustering $C_1 \dot\cup \cdots \dot\cup C_k = \mathcal{P}$, the k-means cost is exactly*

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

Consider the completeness case first.

Completeness. Suppose there exist $S_1, \ldots, S_k \in \binom{[n]}{y}$ such that $\bigcup_{i \in [k]} \mathsf{cov}(S_i, E) = E$. Then, we define a clustering $C_1 \dot\cup \cdots \dot\cup C_k = \mathcal{P}$ as follows: for every $a \in \mathcal{P}$, where $a := \widetilde{A}_E(T)$ for some $T \in E$, let $a \in C_i$ such that $T \in \mathsf{cov}(S_i, E)$ (if there is more than one $i \in [k]$ for which S_i is contained in T then we choose one arbitrarily). We now provide an upper bound on the k-means cost of clustering $\mathcal{C} := \{C_1, \ldots, C_k\}$. (3) implies that for each C_i, for any pair T, T' such that $\widetilde{A}_E(T), \widetilde{A}_E(T') \in C_i$, we have that $||\widetilde{A}_E(T) - \widetilde{A}_E(T')||_2^2 \leq 2\ell$. Hence, if we let W_i be the

k-means cost for the C_i,

$$W_i = \frac{1}{2|C_i|} \sum_{q \in C_i} \sum_{p \in C_i} ||p-q||_2^2 \leq \frac{1}{2|C_i|} \sum_{q \in C_i} \sum_{p \in C_i, p \neq q} 2(z-y)\ell \leq (z-y)|C_i|\ell.$$

Thus, the cost of clustering $\mathcal{C}$ is at most $(z-y)\ell|\mathcal{P}|$. Finally, we turn to the soundness analysis.

Soundness. Consider the optimal k-means clustering $\mathcal{C} := \{C_1, \ldots, C_k\}$ of the instance (i.e., $C_1 \dot\cup \cdots \dot\cup C_k = \mathcal{P}$). We aim at showing that the k-means cost of $\mathcal{C}$ is at least $((z-y) + 2(1-\alpha) - o(1))\ell|\mathcal{P}|$. Given a cluster C_i, let $E_i := \{T \in E : \widetilde{A}_E(T) \in C_i\}$ be the collection of z-sets of E corresponding to C_i. For each $S \in \binom{[n]}{y}$, we define the *degree of S in C_i* to be

$$d_{i,S} := \left|\{T \mid S \subset T \text{ and } \widetilde{A}_E(T) \in C_i\}\right|.$$

Let $t_0 = (2z - 2y - \delta)\ell$, $t_1 = (2z-2y)\ell$, and $t_2 = (2z - 2y + 2 - \delta)\ell$. For each cluster C_i, let

$$F_i = \left|\{(p,q) \in C_i^2 : \|p-q\|_2^2 \geq t_2\}\right|$$
$$M_i = \left|\{(p,q) \in C_i^2 : \|p-q\|_2^2 \in [t_0, t_1]\}\right|$$
$$N_i = \left|\{(p,q) \in C_i^2 : \|p-q\|_2^2 < t_0\}\right|.$$

By (3) and (4), F_i, M_i, and N_i are the number of (ordered) pairs within C_i whose corresponding z-sets in the Johnson Coverage instance intersect in $< y$, $= y$, and $> y$ elements respectively. Let $\Delta_i = \max_{S \in \binom{[n]}{y}} d_{i,S}$. We write the total cost of the clustering as follows.

$$\begin{aligned} & \sum_{i=1}^{k} W_i \\ \geq \quad & \sum_{i=1}^{k} \frac{1}{2|C_i|}\Big(F_i t_2 + M_i t_0\Big) \\ = \quad & \sum_{i=1}^{k} \frac{1}{2|C_i|}\Big((|C_i|^2 - M_i)t_2 + (M_i)t_0 - N_i t_2\Big) \end{aligned} \tag{5}$$

We first upper bound $\sum_i (N_i t_2)/(2|C_i|)$. For each z-set T, there are at most $(z-y)\cdot\binom{z}{z-y-1}\cdot n^{z-y-1}$ sets that intersect with T in at least $y+1$ elements. Therefore, $N_i \leq |C_i| \cdot \max(|C_i|, O(n^{z-y-1}))$ and

$$\sum_{i=1}^{k} \frac{N_i}{2|C_i|} \leq \sum_{i=1}^{k} \max(|C_i|, O(n^{z-y-1})) \leq O(k \cdot n^{z-y-1})).$$

By the definition of Johnson Coverage*, $|E| = \omega(k \cdot n^{z-y-1})$, so $\sum_i N_i t_2/(2|C_i|)$ is at most $o(\ell|\mathcal{P}|)$.

For M_i, we prove the following claim that bounds $M_i/|C_i|$ in terms of Δ_i and $|C_i|$.

Claim 3.18. *For every $i \in [k]$, either $|C_i| = o(|\mathcal{P}|/k)$ or $M_i/|C_i| \leq (1+o(1))\Delta_i + o(|C_i|)$.*

This proves the theorem because we can lower bound (5) as

$$\begin{aligned}
& \sum_{i=1}^{k} \frac{1}{2|C_i|}\Big((|C_i|^2 - M_i)t_2 + (M_i)t_0 - N_i t_2\Big) \\
\geq\ & \sum_{i=1}^{k} \frac{1}{2|C_i|}\Big((|C_i|^2 - M_i)t_2 + (M_i)t_0\Big) - o(\ell|\mathcal{P}|) \\
=\ & \frac{1}{2}\sum_{i=1}^{k}\Big(t_2|C_i| - (t_2 - t_0)(M_i/|C_i|)\Big) - o(\ell|\mathcal{P}|) \\
=\ & \frac{1}{2}\Big(t_2|\mathcal{P}| - (t_2 - t_0)\sum_{i=1}^{k}(M_i/|C_i|)\Big) - o(\ell|\mathcal{P}|) \\
\geq\ & \frac{1}{2}\Big(t_2|\mathcal{P}| - (t_2 - t_0)\sum_{i=1}^{k}\Delta_i\Big) - o(\ell|\mathcal{P}|),
\end{aligned}$$

where the last inequality used the fact that either either $M_i/|C_i| \leq (1+o(1))\Delta + o(|C_i|)$ or $|C_i| = o(|E|/k)$ where we can use trivial bound $M_i/|C_i| \leq |C_i| = o(|\mathcal{P}|/k)$ so that all the terms multiplied by $o(1)$ can be absorbed by the last term $o(\ell|\mathcal{P}|)$. By using the soundness condition $\sum_{i=1}^{k}\Delta_i \leq \alpha|\mathcal{P}|$ and plugging in the values of t_0 and t_2, the total cost is at least by

$$|\mathcal{P}|\frac{t_0}{2} + |\mathcal{P}|\frac{(t_2 - t_0)}{2}\Big(1-\alpha\Big) - o(\ell|\mathcal{P}|) = \ell|\mathcal{P}|\Big((z-y) + (1-\alpha) - \delta/2 - o(1)\Big).$$

Therefore, it remains to proof Claim 3.18.

Proof of Claim 3.18. Note that $M_i \leq \sum_{S\in\binom{[n]}{y}} d_{i,S}^2$. First, consider the set $\mathcal{S} = \{S \in \binom{n}{y} \mid d_{i,S} < \gamma|C_i|\}$ for some small $\gamma > 0$ that will be determined later, and let $E_i^0 = \{T \in E_i \mid \exists S \in \mathcal{S}, S \subseteq T\}$. We have that

$$\frac{1}{|C_i|}\sum_{S\in\mathcal{S}} d_{i,S}^2 \leq \frac{\gamma|C_i|}{|C_i|}\sum_{S\in\mathcal{S}} d_{i,S} \leq \binom{z}{y}\gamma|E_i^0|, \tag{6}$$

because each z-set T can contribute to the degree $d_{i,S}$ of at most $\binom{z}{y}$ different sets.

We then bound $\sum_{S\notin\mathcal{S}} d_{i,S}^2$. Let $\mathcal{S}' = \binom{n}{y} \setminus \mathcal{S}$ and $E_i^1 = \{T \in E_i \mid \exists S_1, S_2 \in \mathcal{S}', S_1 \neq S_2, S_1 \subseteq T, S_2 \subseteq T\}$. We claim that

$$\sum_{S\in\mathcal{S}'} d_{i,S} \leq \binom{z}{y}|E_i^1| + |E_i|. \tag{7}$$

Indeed, consider an element $T \in E_i$, we have that either there is at most one element S of $\mathcal{S}'$ such that $S \subset T$, in which case T will be counted at most once in $\sum_{S\in\mathcal{S}'} d_{i,S}$, or T contains more than one element of $\mathcal{S}'$. The elements of E_i^1 are counted at most $\binom{z}{y}$ times in the sum and from there follows (7).

We claim that if $|E_i^1| > \gamma|E_i|$ or $|E_i|$ is small. From (7) we deduce that there exists $S \in \mathcal{S}'$ such that $d_{i,S} \leq (\binom{z}{y}|E_i^1| + |E_i|)/|\mathcal{S}'|$. Moreover, since $S \in \mathcal{S}'$ we also have that $d_{i,S} \geq \gamma|E_i| \geq \gamma|E_i^1|$ and so, combining the two bounds yields $\gamma|E_i^1| \leq (\binom{z}{y}|E_i^1| + |E_i|)/|\mathcal{S}'|$. Rearranging gives

$$|\mathcal{S}'| \leq \binom{z}{y}/\gamma + \frac{|E_i|}{\gamma|E_i^1|}.$$

If $|E_i^1| > \gamma|E_i|$ we get $|\mathcal{S}'| \leq \binom{z}{y}/\gamma + 1/\gamma^2$. Hence, $|E_i^1|$ is at most $(\binom{y}{z}/\gamma + 1/\gamma^2)^2 \cdot n^{z-y-1}$

since for each pair $S, S' \in \mathcal{S}'$, $S \neq S'$ there is at most n^{z-y-1} $T \in E_i$ such that $S \subset T, S' \subset T$. Thus, $|E_i| \leq (\binom{y}{z}/\gamma + 1/\gamma^2)^2/\gamma \cdot n^{z-y-1}$, proving the claim.

For i with $|E_i^1| \leq \gamma |E_i|$, using (7) together with the above bound on $|E_i^1|$, we obtain

$$\sum_{S \notin \mathcal{S}} d_{i,S}^2 \leq \Delta_i \sum_{S \notin \mathcal{S}} d_{i,S} \leq \Delta_i(|E_i| + \binom{z}{y}|E_i^1|) \leq \Delta_i(1 + \gamma\binom{z}{y})|E_i|. \tag{8}$$

Combining (6) and (8) we deduce that $\frac{1}{|E_i|}\sum_S d_{i,S}^2 \leq (1 + \gamma\binom{z}{y})\Delta_i + \gamma\binom{z}{y}|E_i|$.

Finally, since $|E| = \omega(kn^{z-y-1})$, we can choose $\gamma = o(1)$ such that

$$\gamma\binom{z}{y} = o(1) \text{ and } \left(\left(\binom{y}{z}/\gamma + 1/\gamma^2\right)^2/\gamma\right) \cdot kn^{z-y-1} = o(|E|).$$

Therefore, if $|E_i^1| \leq \gamma|E_i|$, then $M_i/|E_i| \leq (1 + o(1))\Delta_i + o(|E_i|)$, and if $|E_i^1| > \gamma|E_i|$, $|E_i| = o(|E|/k)$. This finishes the proof of the claim and the theorem. □

□

Remark 3.19. *Similar to Remark 3.14, we note here that given $(1-\delta, z, 1)$-Johnson Coverage problem is* NP*-Hard for some constants $\delta > 0$ and $z \in \mathbb{N}$, plugging those parameters in Theorem 3.16, we immediately have that continuous k-*means *is* APX*-Hard in ℓ_2-metric.*

Before, we address continuous case of k-median in ℓ_1-metric, we prove hardness of approximating the continuous case of both k-means and k-median in the Hamming metric. The inapproximability of k-median in ℓ_1-metric then follows by a simple observation.

Theorem 3.20 (k-median without candidate centers in $O(\log n)$ dimensional ℓ_0-metric space)**.** *Assume (α, z, y)-Johnson Coverage* is* NP*-Hard. For every constant $\varepsilon > 0$, 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 $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \{0,1\}^d$ and $\sigma : \mathcal{P} \to \mathcal{C}'$ such that*

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

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

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_0 \geq \left(1 + \frac{(1-\alpha)}{(z-y)} - \varepsilon\right) \cdot \rho n \log n,$$

for some constant $\rho > 0$.

Theorem 3.21 (k-means without candidate centers in $O(\log n)$ dimensional ℓ_0-metric space)**.** *Assume (α, z, y)-Johnson Coverage* is* NP*-Hard. For every constant $\varepsilon > 0$, 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 $\mathcal{C}' := \{c_1, \ldots, c_k\} \subseteq \{0,1\}^d$ and $\sigma : \mathcal{P} \to \mathcal{C}'$ such that*

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

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

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_0^2 \geq \left(1 + \frac{(1-\alpha)(2(z-y)+1)}{(z-y)^2} - \varepsilon\right) \cdot \rho n (\log n)^2,$$

for some constant $\rho > 0$.

Proof of Theorems 3.20 and 3.21. The construction of the point-set $\mathcal{P}$ in the theorem statement is the same as in the proof of Theorem 3.16. The completeness case is the same as in Theorem 3.12. Therefore, we have that the k-means cost of the overall instance is at most $\ell^2 \cdot |\mathcal{P}|$, while the k-median cost is at most $\ell \cdot |\mathcal{P}|$. The soundness analysis to prove the theorem follows by a case analysis, which is elaborated in [CK19] and we skip it here for the sake of brevity. □

We consider below k-median in ℓ_1-metric.

Theorem 3.22 (k-median without candidate centers in $O(\log n)$ dimensional ℓ_1-metric space)**.** *Assume* (α, z, y)*-Johnson Coverage* is* NP*-Hard. For every constant* $\varepsilon > 0$*, 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*

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

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

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_1 \geq \left(1 + \frac{(1-\alpha)}{(z-y)} - \varepsilon\right) \cdot \rho n \log n,$$

for some constant $\rho > 0$.

Proof Sketch. The proof follows from a simple observation that for instances where all the coordinates of the points to be clustered are in $\{0,1\}$, we have that 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 $\{0,1\}$, i.e., we may assume without loss of generality that the ith coordinate of the center is also in $\{0,1\}$. Therefore, simply mimicking the proof of Theorem 3.20 yields the desired theorem statement. See [CK19] for a formal argument. □

Finally, we finish the section by proving the remaining hardness results for continuous k-median and k-means: k-median in ℓ_2-metric and k-means in ℓ_1-metric. Especially, k-median in ℓ_2-metric for $k = 1$ is known as the *geometric median* and is actively studied for bounded d or k [ARR98, BHPI02, KSS10, FL11], but to the best of our knowledge, no (hardness of) approximation algorithms have been studied for general d and k. First, we prove the following lemma that if points are pairwise far apart, there is no good center that is close to all of them.

Lemma 3.23. *For any* $\varepsilon \in (0,1)$*, if* $p_1, \ldots, p_n \in \mathbb{R}^d$ *with* $n > 4/\varepsilon + 1$ *satisfy* $\|p_i - p_j\|_2 \geq \sqrt{2}$*, then, for any* $c \in \mathbb{R}^d$*,* $\max_{i \in [n]} \|c - p_i\|_2 \geq 1 - \varepsilon$.

Proof. Assume towards contradiction that there exists c such that $\max_{i\in[n]} \|c - p_i\|_2 < 1 - \varepsilon$. Without loss of generality let $c = \vec{0}$ and consider the matrix $A \in \mathbb{R}^{n\times n}$ such that $A_{i,j} = \langle p_i, p_j\rangle = (\|p_i\|_2^2 + \|p_j\|_2^2 - \|p_i - p_j\|_2^2)/2 < (2(1-\varepsilon)^2 - 2)/2 \le \varepsilon$. Then A is positive semidefinite, so if $\vec{1}$ denotes the all-ones vector,

$$(\vec{1}^T)A\vec{1} \le \sum_i A_{i,i} + \sum_{i<j} A_{i,j} \le n - \binom{n}{2}\varepsilon \ge 0,$$

which leads to contradiction if $n > 4/\varepsilon + 1$. □

Theorem 3.24 (k-median without candidate centers in $O(\log n)$-dimensional ℓ_2-metric space)**.** *Assume (α, z, y)-Johnson Coverage* is* NP*-Hard. For every constant $\varepsilon > 0$, 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*

$$\sum_{a\in\mathcal{P}} \|a - \sigma(a)\|_2 \le n$$

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

$$\sum_{a\in\mathcal{P}} \|a - \sigma(a)\|_2 \ge \left(\alpha + \sqrt{\frac{z-y+0.5}{z-y}}(1-\alpha) - \varepsilon\right) n$$

Proof. Given an instance (U, E, k) of (α, z, y)-Johnson Coverage*, for any $S \subseteq U$, let $\tau(S)$ be the indicator vector of S. (The dimension can be reduced to $O(\log n)$ by the standard dimension reduction technique.)

In the completeness case, every point is connected to a center at distance $\sqrt{z-y}$. For the soundness case, fix one cluster C and its center $c \in \mathbb{R}^U$. Fix $\varepsilon > 0$, and remove all points from C whose ℓ_2 distance to c is greater than $(\sqrt{z-y+0.5} - \varepsilon)$. Assume $C = \omega(n^{z-y-1})$. We will show that there exists $S' \in \binom{U}{y}$ that covers all but $O(n^{z-y-1})$ sets in C.

We claim that no two y-sets can both cover $\omega(n^{z-y-1})$ number of z-sets in C. Assume towards contradiction that there exist $S', T' \in \binom{U}{y}$ be such that each of them covers at least $t = \omega(n^{z-y-1})$ sets in C. Let $s = |S' \cap T'|$ and $I = S' \cap T'$. Consider the center c and its squared distances to $\tau(S')$ and $\tau(T')$ in coordinates $Q = (S' \cup T') \setminus I$. Since $\tau(S')$ and $\tau(T')$ are different in these coordinates, $(c_i - \tau(S')_i)^2 + (c_i - \tau(T')_i)^2 \ge 0.5$ for each $i \in Q$. Without loss of generality, assume that $(c_i - \tau(S')_i)^2 \ge (y-s)/2 \ge 0.5$.

Now let us consider the points restricted to the coordinates $O = U \setminus (S \cup T)$. First, we claim that we can choose $S_1, \ldots, S_\ell$ with $\ell = \omega(1)$ such that $S_i \in C$, $S' \subseteq S_i$, $S_i \setminus S \subseteq O$, and S_i's are pairwise disjoint. First the number of $S \in C$ that contains S' and intersects T' with in at least one element outside S' is at most $|T'| \cdot n^{z-y-1}$, which is only an $o(1)$ fraction of the z-sets in C covered by S'. Remove them from C. Now, consider a greedy iteration for $i = 1, \ldots, \ell$ where we pick an arbitrary set $S_i \in C$ that contains S', and remove all sets from C that insect S_i outside S'. Each time, the number of removed sets is only $(z-y)n^{z-y-1}$, and since $C = \omega(n^{z-y-1})$, this process can continue $\ell = o(1)$ iterations.

Then $\tau(S_i)$ and $\tau(S_j)$ on coordinates in O are at distance at least $\sqrt{2(z-y)}$ from each other. Since the contribution of $\|c - S_i\|_2^2$ from Q is already at least 0.5, Lemma 3.23 shows

there is no center can cover every point in C at distance at most $\sqrt{z-y+0.5}-o(1)$, leading to the desired contradiction.

Now we prove that there exist a bounded number of y-sets that collectively cover every z-sets. Consider the following process starting with $C' = C$ and $i = 1$.

1. Let S_i be an arbitrary set from C'.
2. Delete all $S \in C'$ such that $|S \cap S_i| \geq y$.
3. If C' is nonempty, increase i by 1 and repeat from 1.

Let t be the final value of i, and consider $S_1, \ldots, S_t$. They are at distance at least $\sqrt{2(z-y+1)}$ from each other, so again by Lemma 3.23 and using that $\sqrt{2(z-y+1)}/\sqrt{2} = \sqrt{z-y+1} > \sqrt{z-y+0.5}$, t is most some absolute constant. Then y-sets in $\{S' \subseteq U : |S'| = y \text{ and } S' \subseteq S_i \text{ for some } i \in [t]\}$ cover all sets in C. A similar argument for $C^* = \{S : \|\tau(S) - c\|_2 < \sqrt{z-y} - \varepsilon\}$ and $y+1$ implies that $|C^*| \leq O(n^{z-y-1})$.

So far, we showed that (1) there are a constant number of y-sets that collectively cover every $S \in C$ (2) except one set, every y-set can cover at most $O(n^{z-y-1})$ sets in C, and (3) $|C^*| \leq O(n^{z-y-1})$. Therefore, whenever $|C| = \omega(n^{z-y-1})$, there exists $S' \in \binom{U}{y}$ that covers all but a subconstant fraction of sets in C. Since $|E| = \omega(kn^{y-z-1})$, we can use an argument similar to Theorem 3.16 to show that in the soundness case, the fraction of points that are covered by a center at distance at most $\sqrt{z-y+0.5}-\varepsilon$ is at most $\alpha + o(1)$, and at most an $o(1)$ fraction of points are covered at distance at most $\sqrt{z-y}-\varepsilon$. This proves the theorem. □

Theorem 3.25 (k-means without candidate centers in poly(n)-dimensional ℓ_1-metric space)**.** *Assume (α, z, y)-Johnson Coverage* is* NP*-Hard. For every constant $\varepsilon > 0$, 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*

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_1^2 \leq n$$

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

$$\sum_{a \in \mathcal{P}} \|a - \sigma(a)\|_1^2 \geq \left(\alpha + (1-\alpha)\frac{(z-y+1)^2}{(z-y)^2} - \varepsilon\right) n.$$

Proof. Given an instance (U, E, k) of (α, z, y)-Johnson Coverage*, for any $S \subseteq U$, let $\tau(S)$ be the indicator vector of S. We analyze the completeness and soundness of this simple reduction. Since every point is in the boolean hypercube, the embedding of 3.12 ensures that one can reduce the dimension to $O(\log n)$ with losing an arbitrarily small constant in the inapproximability factor.

In the completeness case, every point is connected to a center at distance $z - y$. For the soundness case, fix one cluster C and its center $c \in \mathbb{R}^U$. Fix $\varepsilon > 0$, and remove all points from C whose ℓ_1 distance to c is at least $(z - y + 1 - \varepsilon)$.

First, note that for every $S, T \in C$, $|S \cap T| \geq y$. Otherwise, $\|\tau(S) - \tau(T)\|_1 \geq 2(z-y+1)$, so there is no center that are at distance strictly less than $z - y + 1$ from both $\tau(S)$ and $\tau(T)$.

Fix any $S \in C$. Then every $T \in C$ is covered by a y-set in $\mathcal{F} = \{S' \subseteq U : |S'| = y \text{ and } S' \subseteq S\}$. Furthermore, argument similar to the proof of Theorem 3.24 show except possibly one, every y-set can cover at most $O(n^{z-y-1})$ sets in C, and the number of points that are covered at distance strictly less than $z - y$ is also at most $O(n^{z-y-1})$.

Therefore, whenever $|C| = \omega(n^{z-y-1})$, there exists $S' \in \binom{U}{y}$ that covers all but a sub-constant fraction of sets in C. Since $|E| = \omega(kn^{y-z-1})$, we can use an argument similar to Theorem 3.16 to show that in the soundness case, the fraction of points that are covered by a center at distance at most $z - y + 1 - \varepsilon$ is at most $\alpha + o(1)$, and at most an $o(1)$ fraction of points are covered at distance at most $z - y - \varepsilon$. This proves the theorem. □

3.5 Integrality Gaps

In this subsection we present improved integrality gaps. Given an instance of discrete k-median or k-means with a set of candidate centers $\mathcal{C}$, a set of points $\mathcal{P}$, and distances $\{d_{p,c}\}_{p\in\mathcal{P},c\in\mathcal{C}}$, the basic LP relaxation for k-median or k-means is the following. (For k-means, $d_{p,c}$ becomes a squared distance.)

$$
\begin{aligned}
\text{Minimize} \quad & \sum_{p\in\mathcal{P}} \sum_{c\in\mathcal{C}} x_{p,c} \cdot d_{p,c} \\
\text{Subject to} \quad & \sum_{c\in C} x_{p,c} = 1 && \forall p \in \mathcal{P} \\
& x_{p,c} \le y_c && \forall p \in \mathcal{P}, c \in \mathcal{C} \\
& \sum_{c\in C} y_c \le k \\
& x, y \ge 0.
\end{aligned}
$$

The basic SDP relaxation, which replaces $x_{p,c}$ and y_c by $\|v_{p,c}\|_2^2$ and $\|u_c\|_2^2$ for some vectors $\{v_{p,c}\}$ and $\{u_c\}$ with additional constraints, is the following.

$$
\begin{aligned}
\text{Minimize} \quad & \sum_{p\in\mathcal{P}} \sum_{c\in\mathcal{C}} \|v_{p,c}\|_2^2 \cdot d_{p,c} \\
\text{Subject to} \quad & \langle v_0, v_0 \rangle = 1 \\
& \langle v_{p,c}, v_0 \rangle = \|v_{p,c}\|_2^2 && \forall p \in \mathcal{P}, c \in \mathcal{C} && (9) \\
& \langle u_c, v_0 \rangle = \|u_c\|_2^2 && \forall c \in \mathcal{C} && (10) \\
& \langle v_{p,c}, u_c \rangle = \|v_{p,c}\|_2^2 && \forall p \in \mathcal{P}, c \in \mathcal{C} && (11) \\
& \| \sum_{c\in C} v_{p,c} - v_0 \|_2^2 = 0 && \forall p \in \mathcal{P} && (12) \\
& \sum_{c\in C} \|y_c\|_2^2 \le k
\end{aligned}
$$

We give stronger integrality gap instances than the standard notion. The LP or SDP relaxation has a *robust* gap of $\alpha > 1$ if there exists a family of instances for infinitely many values of k where for any $\varepsilon > 0$, there exists $\delta > 0$ and $k_0 \in \mathbb{N}$ such that for all instances in the family with $k \ge k_0$, the gap between the optimal fractional solution and the optimal integral solution

is at least $\alpha - \varepsilon$ even when the optimal integral solution is allowed to open $k + \delta k$ facilities.

Our instances additionally satisfy *well-separated* the property that every pair of points in $\binom{\mathcal{P}\cup\mathcal{C}}{2}$ are at distance at least 1 far part, where each instance is normalized so that 1 is the average connection cost of each point in the SDP solution.

The basic LP relaxation for metric k-median has a gap of 2 under the standard notion [JMS02], but this instance does not give a robust gap. The best approximation algorithms for k-median and k-means in both general and Euclidean metrics bound the robust gap of the LP relaxation [BPR+15, ANSW20], and to the best of our knowledge, for well-separated instances, no robust gap of the LP relaxation better than computational hardness results [GK99, CK19] were previously known.

Theorem 3.26. *Fix any $\varepsilon > 0$. For discrete k-median in ℓ_1 and discrete k-means in ℓ_2, there is a family of well-separated instances where the SDP relaxation has a gap of at least $149/125 - \varepsilon \approx 1.192 - \varepsilon$, even when the integral solution opens $\Omega(k)$ more centers.*

Proof. Consider the complete graph $K_n = ([n], E)$ with $E = \binom{[n]}{2}$. Let $\tau : 2^E \to \mathbb{R}^n$ where for any $p \subseteq E$, $\tau(p)$ is defined to be the characteristic vector of $\cup_{e\in p} e$. The set points $\mathcal{P}$ is defined to be $\mathcal{P} := \{\tau(p) : p \text{ forms a 4-clique in } K_n\}$. The set C of candidate centers, for each edge $e \in E$, has a center $\tau(e)$. To simplify notation, we use p (resp. e) to also denote $\tau(p)$ (resp. $\tau(e)$) when p (resp. e) is used as part of the clustering instance.

We say that an edge $e \in E$ *covers* a 4-clique p if $e \in p$; in this case a center $e \in \mathcal{C}$ also *covers* a point $p \in \mathcal{P}$. Note that $\|e - p\|_0 = \|e - p\|_1 = 2$ if e covers p and at least 4 otherwise. Since each 4-clique has 6 edges and each edge yields m centers, note that each point $p \in \mathcal{P}$ is covered by exactly 6 centers. Therefore, there is an LP solution that picks $1/6$ of each $e \in C$ and fractionally covers every $p \in \mathcal{P}$. We prove that even for the SDP, there is such a solution that picks $1/5$ of each $e \in C$.

Claim 3.27. *There exists a feasible SDP solution where every $p \in \mathcal{P}$ is fractionally connected to only centers at distance at most 2, and $\|u_e\|_2^2 = 1/5$ for every e.*

Proof. Let $t = 5$. Let v_0 and $\{w_e, w'_e\}_{e\in E}$ be pairwise orthogonal unit vectors. We explicitly construct the vectors for the SDP relaxation.

- For each $e \in E$, $u_e = \frac{v_0}{t} + \left(\frac{(t-1)\sqrt{t+1}}{t^2}\right) w_e + \left(\frac{\sqrt{t-1}}{t^2}\right) w'_e$.
- For each $e \in E$, and a 4-clique p, if $e \in p$,

$$v_{p,e} = \frac{1}{t+1} \cdot v_0 + \frac{t}{(t+1)^{3/2}} \cdot w_e - \sum_{f\in p\setminus\{e\}} \frac{1}{(t+1)^{3/2}} \cdot w_f.$$

  Otherwise $v_{p,e} = 0$.

Note that every point $p \in \mathcal{P}$ is only connected to e that covers p. We check each constraint of the SDP. The first constraint $\langle v_0, v_0\rangle = 1$ is true by definition. Since

$$\|u_e\|_2^2 = \frac{1}{t^2} + \frac{(t-1)^2(t+1)}{t^4} + \frac{t-1}{t^4} = \frac{1}{t^2} + \frac{t-1}{t^2} = \frac{1}{t} = \langle v_0, u_e\rangle,$$

it satisfies (10). Since for $e \in p$,

$$\|v_{p,e}\|_2^2 = \frac{1}{(t+1)^2} + \frac{t^2}{(t+1)^3} + \frac{t}{(t+1)^3} = \frac{1}{t+1} = \langle v_0, v_{p,e} \rangle,$$

it satisfies (9). Furthermore, for each $p \in \mathcal{P}$, $\sum_{e \in p} v_{p,e} = v_0$, since the coefficient of w_e in the sum for any $e \in p$ is $\frac{t}{(t+1)^{3/2}}$ (from $v_{p,e}$) minus t times $\frac{1}{(t+1)^{3/2}}$ (from every other $v_{p,f}$), which is 0. It satisfies (12). Finally, (11) can be checked as for every 4-clique p, $e \in p$,

$$\langle u_e, v_{p,e} \rangle = \frac{1}{t} \cdot \frac{1}{t+1} + \frac{(t-1)\sqrt{t+1}}{t^2} \cdot \frac{t}{(t+1)^{3/2}} = \frac{1}{t(t+1)} + \frac{t-1}{t(t+1)} = \frac{1}{t+1} = \|v_{p,e}\|_2^2.$$

Since the solution satisfies every SDP constraint, each $p \in \mathcal{P}$ is only connected to a center covering p, and $\|u_e\|_2^2 = 1/t$ for every e, the claim is proved. □

Therefore the optimal relaxation value is at most $2 \cdot \binom{n}{4}$ when $k = \binom{n}{2}/5$. Now we consider if we pick at most k edges E integrally, how many 4-cliques can be covered by them. Equivalently, we ask if we pick at least $\binom{n}{2} - k$ edges, how many 4-cliques are completely contained by them. The clique density theorem of Reiher [Rei16] answers this question, proving that if we pick at least an $(1 - 1/t)$ fraction of edges for some integer $t \geq 4$, the number of 4-cliques completely contained in them is at least that of complete t-partite graph with each partition having the same size. Note that in the complete t-partite graph, the probability that a random 4-tuple becomes a clique is roughly $1 \cdot \frac{t-1}{t} \cdot \frac{t-2}{t} \cdot \frac{t-3}{t}$.

Theorem 3.28 ([Rei16]). *Let $t \geq 4$ be an integer. In every graph on n vertices with at least $\frac{(t-1)}{t} \cdot \binom{n}{2}$ edges, the number of 4-cliques is at least*

$$\binom{n}{4} \cdot \frac{t(t-1)(t-2)(t-3)}{t^4}.$$

Applying the above theorem with $t = 5$ shows that if we pick k edges, the fraction of 4-cliques covered by them is at most $1 - (5 \cdot 4 \cdot 3 \cdot 2)/5^4 = 24/125$. Therefore, while every point is connected at distance 2 the SDP solution, in the integral solution at least a $24/125$ fraction of the points are connected at distance at least 4. The gap is at least $(2 + 2(24/125)/2) = 149/125 \approx 1.192$. □

4 NP-Hardness of Approximatig 3-Hypergraph Vertex Coverage problem

In this section, we prove the following theorem showing that for any $\varepsilon > 0$, $(7/8 + \varepsilon, 3, 1)$-Johnson Coverage Problem is NP-hard for randomized reductions (even in the dense case). By the results in Sections 3.3 and 3.4, we obtain the inapproximability results for clustering problems stated in Theorem 1.5 and Table 1.

Theorem 4.1. *For any $\varepsilon > 0$, given a simple 3-hypergraph $\mathcal{H} = (V, H)$ with $n = |V|$, it is NP-hard to distinguish between the following two cases:*

- **Completeness:** *There exists $S \subseteq V$ with $|S| = n/2$ that intersects every hyperedge.*

- **Soundness:** *Any subset $S \subseteq V$ with $|S| \leq n/2$ intersects at most a $(7/8 + \varepsilon)$ fraction of hyperedges.*

Furthermore, with randomized reductions, the above hardness holds when $|H| = \omega(n^2)$.

We first prove Theorem 4.1 without the lower bound on $|H|$; Section 4.1 presents the construction and proves the completeness of the reduction, and Section 4.2, 4.3, and 4.4 analyze its soundness. Finally, Section 4.5 shows how to ensure $|H| = \omega(n^2)$.

4.1 Overview and Construction

Consider the setting of Theorem 4.1. A random set of $|V|/2$ elements will intersect each hyperedge with probability $7/8$, so the theorem says that it is hard to do even slightly better than the random solution. It is similar to the notion of *approximate resistance* mainly studied for constraint satisfaction problems. Indeed, given an instance of Max 3-SAT with variables $\{x_1, \ldots, x_n\}$ and clauses $\{(\ell_{i,1} \cup \ell_{i,2} \cup \ell_{i,3})\}_{i \in [m]}$ where each $\ell_{i,j}$ denotes a literal x_k or $\overline{x_k}$, consider the simple reduction of creating a vertex for each literal and a hyperedge $(\ell_{i,1}, \ell_{i,2}, \ell_{i,3})$ for each clause. This reduction almost proves the above theorem, except that the soundness property only holds for S that satisfies $|S \cap \{x_i, \overline{x_i}\}| \leq 1$ for each $i \in [n]$. However, the resulting hypergraph produced by this reduction combined with Håstad's celebrated hardness for Max 3-SAT [Hås01] is always bipartite due to the underlying bipartite structure of the outer verifier, so there is a vertex cover of size at most $|V|/2$ even in the soundness case.

We bypass the above problem by plugging in Håstad's inner verifier to the outer verifier constructed in Guruswami et al. [DGKR05] and Khot [Kho02]. This outer verifier is called a *multilayered PCP* and used for proving hardness of the covering version of our problem called k-Hypergraph Vertex Cover. We give a new analysis for the multilayered PCP and combine the analysis for Håstad's inner verifier to bound the number of uncovered hyperedges for all S with $|S| \leq |V|/2$.

We now formally present the reduction. We first describe multilayered PCPs that we use.

Definition 4.2. *An ℓ-layered PCP $\mathcal{M}$ consists of*

- *An ℓ-partite graph $G = (V, E)$ where $V = \cup_{i=1}^{\ell} V_i$. Let $E_{i,j} = E \cap (V_i \times V_j)$.*
- *Sets of alphabets $\Sigma_1, \ldots, \Sigma_\ell$.*
- *For each edge $e = (v_i, v_j) \in E_{i,j}$, a surjective projection $\pi_e : \Sigma_j \to \Sigma_i$.*

Given an assignment $(\sigma_i : V_i \to \Sigma_i)_{i \in [\ell]}$, an edge $e = (v_i, v_j) \in E_{i,j}$ is satisfied *if $\pi_e(\sigma_j(v_j)) = \sigma_i(v_i)$. There are additional properties that $\mathcal{M}$ can satisfy.*

- *η-smoothness: For any $i < j$, $v_j \in V$, and $x, y \in \Sigma_j$, $\Pr_{(v_i,v_j) \in E_{i,j}}[\pi_{(v_i,v_j)}(x) = \pi_{(v_i,v_j)}(y)] \leq \eta$.*
- *Path-regularity: Call a sequence $p = (v_1, \ldots, v_\ell)$* full path *if $(v_i, v_{i+1}) \in E_{i,i+1}$ for every $1 \leq i < \ell$, and let $\mathcal{P}$ be the distribution of full paths obtained by (1) sampling a random vertex $v_1 \in V_1$ and (2) for $i = 2, \ldots, \ell$, sampling v_i from the neighbors of v_{i-1} in $E_{i-1,i}$. $\mathcal{M}$ is called* path-regular *if for any $i < j$, sampling $p = (v_1, \ldots, v_\ell)$ from $\mathcal{P}$ and taking (v_i, v_j) is the same as sampling uniformly at random from $E_{i,j}$.*

Theorem 4.3 ([DGKR05, Kho02]). *For any $\tau, \eta > 0$ and $\ell \in \mathbb{N}$, given an ℓ-layered PCP $\mathcal{M}$ with η-smoothness and path-regularity, it is NP-hard to distinguish between the following cases.*

- **Completeness:** *There exists an assignment that satisfies every edge $e \in E$.*
- **Soundness:** *For any $i < j$, no assignment can satisfy more than an τ fraction of edges in $E_{i,j}$.*

Given an ℓ-layered PCP $\mathcal{M}$ described as above, we construct the reduction from to Johnson Coverage. For simplicity of presentation, the produced instance will be vertex-weighted and edge-weighted, so that the problem becomes "choose a set of vertices of total weight at most k to maximize the total weight of covered edges." Vertex weights can be easily removed by duplicating vertices according to weights and creating edges between duplicated vertices with appropriate weights. Edge weights will be handled in Section 4.5.

- Let $C_i := \{\pm 1\}^{\Sigma_i}$ and $U_i := V_i \times C_i$. The resulting hypergraph will be denoted by $\mathcal{H} = (U, H)$ where $U = \cup_{i=1}^{\ell}(V_i \times C_i)$. The weight of vertex $(v, x) \in V_i \times C_i$ is

$$w(v, x) := \frac{1}{\ell} \cdot \frac{1}{|V_i|} \cdot \frac{1}{|C_i|}.$$

  Note that the sum of all vertex weights is 1.

- Let $\mathcal{D}_I$ be the distribution where $i \in [\ell]$ is sampled with probability $(\ell - i)^2/(6\ell(\ell - 1)(2\ell - 1))$, and $\mathcal{D}$ be the distribution over $(i, j) \in [\ell]^2$ where i is sampled from $\mathcal{D}_I$ and j is sampled uniformly from $\{i + 1, \ldots, \ell\}$. For each $i < j$, we create a set of hyperedges $H_{i,j}$ that have one vertex in U_i and two vertices in U_j. Fix each $e = (v_i, v_j) \in E_{i,j}$ and a set of three vertices $t \subseteq (\{v_i\} \times C_i) \cup (\{v_j\} \times C_j)$. The weight $w(t)$ is (the probability that (i, j) is sampled from $\mathcal{D}$) $\cdot$ $(1/|E_{i,j}|)$ $\cdot$ (the probability that t is sampled from the following procedure). The reduction is parameterized by $\delta > 0$ determined later.
  - For each $a \in \Sigma_i$, sample $x_a \in \{\pm 1\}$.
  - For each $b \in \Sigma_j$,
    - Sample $y_b \in \{\pm 1\}$.
    - If $x_{\pi(b)} = -1$, let $z_b = y_b$ with probability $1 - \delta$ and $z_b = -y_b$ otherwise.
    - If $x_{\pi(b)} = 1$, let $z_b = -y_b$.
  - Output $\{(v_i, x), (v_j, y), (v_j, z)\}$.

  Note that the sum of all hyperedge weights is also 1.

Completeness. If $\mathcal{M}$ admits an assignment $(\sigma_i : V_i \to \Sigma_i)_{i \in [\ell]}$ that satisfies every edge $e \in E$, let $S := \{(v_i, x) : v_i \in V_i, x_{\sigma_i(v_i)} = -1\}$. Fix any $e = (v_i, v_j) \in E_{i,j}$ and consider the above sampling procedure to sample $x \in \{\pm 1\}^{\Sigma_i}$ and $y \in \{\pm 1\}^{\Sigma_j}$ when $b = \sigma_j(v_j)$. Since $\pi_e(\sigma_j(v_j)) = \sigma_i(v_i)$, at least one of $x_{\sigma_i(v_i)}$, $y_{\sigma_j(v_j)}$, $z_{\sigma_j(v_j)}$ must be -1 always. This proves that S intersects every hyperedge with nonzero weight.

4.2 Soundness.

In the soundness case, we want to prove that any subset of weight at most $1/2$ intersects hyperedges of total weight at most $7/8 + o(1)$. We prove the equivalent statement that for any

$S \subseteq V$ of weight *greater than* $1/2$ contains hyperedges of total weight approximately *at least* $1/8 - o(1)$.

Fix a set $S \subseteq V$ with $w(S) \geq 1/2$. Let $S_i := S \cap V_i$. Let $F = \{e \in H : e \in S\}$ and $F_{i,j} := F \cap H_{i,j}$. Our goal is to show $w(F)$ is approximately at least $1/8 - o(1)$.

Given a vertex $v \in V_i$, let $C_v := \{v\} \times C_i \subseteq U$ and

$$\alpha_v := \ell |V_i| \cdot \left(\sum_{v \in (S \cap C_v)} w(v) \right)$$

be the normalized weight of S in C_v. Given vertices $v_i \in V_i$ and $v_j \in V_j$ with $i < j$, let $D_{i,j} = \Pr_{(i',j') \sim \mathcal{D}}[i = i', j = j']$ and

$$\beta_{v_i,v_j} := \frac{1}{D_{i,j}} \cdot |E_{i,j}| \cdot \left(\sum_{e \in (F \cap H(C_{v_i} \cup C_{v_j}))} w(e) \right)$$

be the normalized weight of F in $H(C_{v_i} \cup C_{v_j})$, where given $T \subseteq V$, $H(T)$ is defined as $\{e \in H : e \subseteq T\}$. Note that all $\alpha_v, \beta_{v_i,v_j}$ are in $[0,1]$. Furthermore,

$$\mathbb{E}_{i \in [\ell]} \mathbb{E}_{v \in V_i} [\alpha_v] = w(S),$$

and

$$\mathbb{E}_{(i,j) \sim \mathcal{D}} \mathbb{E}_{(v_i,v_j) \in E_{i,j}} [\beta_{v_i,v_j}] = w(F).$$

Let $\alpha_i := \mathbb{E}_{v \in V_i}[\alpha_v]$ and $\beta_{i,j} := \mathbb{E}_{(v_i,v_j) \in E_{i,j}}[\beta_{v_i,v_j}]$. Finally, for each full path $p = (v_1, \ldots, v_\ell)$, let $\alpha_{p,i} := \alpha_{v_i}$ and $\beta_{p,i,j} := \beta_{v_i,v_j}$.

Call a triple (p,i,j) *good* if $\beta_{p,i,j} < \alpha_{p,i}\alpha_{p,j}^2 - \varepsilon$. The following lemma says that we are done if few triples are good.

Lemma 4.4. *If* $\Pr_{p \in \mathcal{P}, (i,j) \in \mathcal{D}}[(p,i,j) \text{ is good}] \leq \varepsilon$, *then* $w(F) \geq 1/8 - 3(\sqrt{\varepsilon} + 1/\ell)$.

Therefore, it remains to consider the case that the condition of Lemma 4.4 does not hold. Then, there exists $i < j$ such that $\Pr_{p \in \mathcal{P}}[(p,i,j) \text{ is good}] \geq \varepsilon$. The following lemma, essentially from Håstad's analysis for Max 3-SAT [Hås01], states that it cannot happen if there is no good assignment for the multilayered PCP instance $\mathcal{M}$. For completeness, we reproduce a proof in Section 4.4

Lemma 4.5. *Fix any* $i < j$. *If* $\Pr_{p \in \mathcal{P}}[(p,i,j) \text{ is good}] \geq \varepsilon$, *then there is an assignment for* $\mathcal{M}$ *that satisfies at least an* $(\varepsilon\delta^4/2) \cdot (\varepsilon/2 - 2\delta - 2\eta\delta^{-4})^2$ *fraction of edges in* $E_{i,j}$.

First take small enough $\varepsilon > 0$ and large enough $\ell \in \mathbb{N}$ will ensure $w(F) \geq 1/8 - 3(\sqrt{\varepsilon} + 1/\ell)$ is almost at least $1/8$. Choosing $\delta = \varepsilon^2/4$, $\eta = \delta^6$ will ensure that the guarantee in Lemma 4.5 is at least ε^c for some absolute constant $c \in \mathbb{N}$. By taking τ in Theorem 4.3 smaller than that, we can ensure that $\Pr_{p \in \mathcal{P}}[(p,i,j) \text{ is good}] \leq \varepsilon$ and $w(F) \geq 1/8 - 3(\sqrt{\varepsilon} + 1/\ell)$ in the soundness case, proving Theorem 4.1.

4.3 Proof of Lemma 4.4

Proof. Recall that for any $i < j$, sampling $p = (v_1, \ldots, v_\ell) \in \mathcal{P}$ and choosing (v_i, v_j) is the same as sampling $(v_i, v_j) \in E_{i,j}$ uniformly at random. Therefore,

$$\begin{aligned} w(F) &= \mathbb{E}_{(i,j)\sim\mathcal{D}}\mathbb{E}_{(v_i,v_j)\in E_{i,j}}[\beta_{v_i,v_j}] \\ &= \mathbb{E}_{(i,j)\sim\mathcal{D}}\mathbb{E}_{p\in\mathcal{P}}[\beta_{p,i,j}] \\ &= \mathbb{E}_{p\in\mathcal{P}}\mathbb{E}_{(i,j)\sim\mathcal{D}}[\beta_{p,i,j}]. \end{aligned}$$

Say p is *atypical* $\Pr_{(i,j)\in\mathcal{D}}[(p,i,j) \text{ is good}] \geq \sqrt{\varepsilon}$. Since $\Pr_{p\in\mathcal{P},(i,j)\in\mathcal{D}}[(p,i,j) \text{ is good}] \leq \varepsilon$,

$$\Pr_{p\in\mathcal{P}}[p \text{ is atypical}] \leq \sqrt{\varepsilon}.$$

Fix a typical p. Then

$$\mathbb{E}_{(i,j)\sim\mathcal{D}}[\beta_{p,i,j}] \geq \mathbb{E}_{(i,j)\sim\mathcal{D}}[\alpha_{i,p}\alpha_{j,p}^2 - \varepsilon] - \sqrt{\varepsilon},$$

since we can apply the lower bound $\beta_{p,i,j}$ by $\alpha_{i,p}\alpha_{j,p}^2 - \varepsilon$ with probability at least $1 - \sqrt{\varepsilon}$ and 0 otherwise.

Now we analyze $\mathbb{E}_{(i,j)\sim\mathcal{D}}[\alpha_{i,p}\alpha_{j,p}^2]$. Recalling the definition of $\mathcal{D}$ and applying Cauchy-Schwarz,

$$\mathbb{E}_{(i,j)\sim\mathcal{D}}[\alpha_{i,p}\alpha_{j,p}^2] \geq \mathbb{E}_{i\sim\mathcal{D}_I,j\in\{i+1,\ldots,\ell\}}[\alpha_{i,p}\alpha_{j,p}^2] \geq \mathbb{E}_{i\sim\mathcal{D}_I,j,k\in\{i+1,\ldots,\ell\}}[\alpha_{i,p}\alpha_{j,p}\alpha_{k,p}]. \tag{13}$$

We compare the RHS of (13) to

$$\big(\mathbb{E}_i[\alpha_{i,p}]\big)^3 = \mathbb{E}_{i,j,k\in[l]}[\alpha_{i,p}\alpha_{j,p}\alpha_{k,p}]. \tag{14}$$

If we fix $i < j < k$, the probability that the monomial $\alpha_{i,p}\alpha_{j,p}\alpha_{k,p}$ contributes to the expectation, after incorporating permutations between i, j, k, is $6/\ell^3$ in (14) and

$$\frac{(\ell-i)^2}{\ell(\ell-1)(2\ell-1)/6} \cdot \frac{2}{(\ell-i)^2} = \frac{2\cdot 6}{\ell(\ell-1)(2\ell-1)}$$

in (13), which is greater than $6/\ell^3$. Since $\Pr[i = j \text{ or } j = k \text{ or } k = i] \leq \frac{3}{\ell}$ when we sample $i, j, k \in [\ell]$ independently,

$$\mathbb{E}_{i\sim\mathcal{D}_I,j,k\in\{i+1,\ldots,\ell\}}[\alpha_{i,p}\alpha_{j,p}\alpha_{k,p}] \geq \mathbb{E}_{i,j,k\in[l]}[\alpha_{i,p}\alpha_{j,p}\alpha_{k,p}] - \frac{3}{\ell} = \big(\mathbb{E}_i[\alpha_{i,p}]\big)^3 - \frac{3}{\ell}.$$

Therefore, if p is typical, then

$$\mathbb{E}_{(i,j)\sim\mathcal{D}}[\beta_{p,i,j}] \geq \left(\mathbb{E}_{i\in[\ell]}[\alpha_i]\right)^3 - \frac{3}{\ell} - 2\sqrt{\varepsilon}.$$

Using $\mathbb{E}_p \mathbb{E}_i[\alpha_{p,i}] \geq 1/2$ and Jensen's inequality,

$$\begin{aligned} w(F) &= \mathbb{E}_{p\in\mathcal{P}} \mathbb{E}_{(i,j)\sim\mathcal{D}}[\beta_{p,i,j}] \\ &\geq \mathbb{E}_{p\in\mathcal{P}} \left[\left(\mathbb{E}_{i\in[\ell]}[\alpha_i] \right)^3 \right] - \frac{3}{\ell} - 2\sqrt{\varepsilon} - \Pr_{p\in\mathcal{P}}[p \text{ is atypical}] \\ &\geq \left(\mathbb{E}_{p\in\mathcal{P}} \mathbb{E}_{i\in[\ell]}[\alpha_i] \right)^3 - \frac{3}{\ell} - 3\sqrt{\varepsilon} \\ &\geq \frac{1}{8} - \frac{3}{\ell} - 3\sqrt{\varepsilon}. \end{aligned}$$

□

4.4 Proof of Lemma 4.5

Proof. Fix $i < j$ given in the condition of the lemma. Call an edge $e = (v_i, v_j) \in E_{i,j}$ good if $\beta_{v_i,v_j} < \alpha_{v_i}\alpha_{v_j}^2 - \varepsilon$. For $v_i \in V_i$, let $f_{v_i} : C_i \to \{0,1\}$ be the indicator function of $S \cap (\{v_i\} \times C_i)$, and Let $v_j \in V_j$, let $g_{v_j} : C_j \to \{0,1\}$ be the indicator function of $S \cap (\{v_j\} \times C_j)$.

The path regularity and the promise of the lemma implies that at least an ε fraction of $e \in E_{i,j}$ is good. Fix such an edge $e = (v_i, v_j)$. For notational simplicity, let $\pi = \pi_e$, $L = \Sigma_i$, $R = \Sigma_j$, $f := f_{v_i}$ and $g := g_{v_j}$.

We use the standard notations in analysis of boolean functions. See [Hås01, O'D14] for references. For two functions $f_1, f_2 : C_i \to \mathbb{R}$, let $\langle f_1, f_2 \rangle := \mathbb{E}_{x\in\{\pm1\}^L}[f_1(x)f_2(x)]$ be the inner product between f_1 and f_2. For $A \subseteq L$, let $\chi_A : C_i \to \mathbb{R}$ defined as $\chi_A(x) = \prod_{a\in A} x_a$. It is well known that $\{\chi_A\}_{A\subseteq L}$ forms an orthonormal basis, so that f can be written as $\sum_{A\subseteq L} \widehat{f}(A)\chi_A$ with $\widehat{f}(A) = \langle f, \chi_A \rangle$. Define $\{\chi_B\}_{B\in\subseteq R}$ similarly and write g as $\sum_{B\subseteq R} \widehat{g}(B)\chi_B$.

Now note that $\alpha_i = \mathbb{E}_x[f(x)] = \widehat{f}(\emptyset)$, $\alpha_j = \mathbb{E}_y[g(y)] = \widehat{g}(\emptyset)$, and $\beta_{i,j} = \mathbb{E}_{x,y,z}[f(x)g(y)g(z)]$ where x, y, z were jointly sampled the reduction given e. Expanding Fourier decompositions for f and g,

$$\begin{aligned} \mathbb{E}_{x,y,z}[f(x)g(y)g(z)] &= \sum_{A\subseteq L, B\subseteq R, C\subseteq R} \widehat{f}(A)\widehat{g}(B)\widehat{g}(C)\mathbb{E}[\chi_A(x)\chi_B(y)\chi_C(z)] \\ &= \sum_{B\subseteq R, A\subseteq\pi(B)} \widehat{f}(A)\widehat{g}(B)^2\mathbb{E}[\chi_A(x)\chi_B(y)\chi_B(z)] \\ &= \sum_{B\subseteq R} \widehat{g}(B)^2 \sum_{A\subseteq\pi(B)} \widehat{f}(A)\mathbb{E}[\chi_A(x)\chi_B(y)\chi_B(z)]. \end{aligned} \tag{15}$$

The second equality holds because if $b \in B \setminus C$, then $\chi_B(y)$ contains y_b and it is independent from any other variable appearing in $\chi_A(x)\chi_B(y)\chi_C(z)$. Similarly, the existence of $c \in C \setminus B$ or $a \in A \setminus (B \cup C)$ will make $\chi_A(x)\chi_B(y)\chi_C(z)$ vanish.

Suppose $B \subseteq \pi^{-1}(a)$ for some $a \in L$. For each $b \in B$, $\mathbb{E}[y_b z_b] = -1$ if $x_a = 1$ and $(1-2\delta)$ otherwise, so

$$\mathbb{E}[\chi_B(y)\chi_B(z)] = \frac{1}{2}\big((-1)^{|B|} + (1-2\delta)^{|B|}\big),$$

and

$$\mathbb{E}[x_a\chi_B(y)\chi_B(z)] = \frac{1}{2}\big((-1)^{|B|} - (1-2\delta)^{|B|}\big).$$

Therefore, if we consider $\mathbb{E}[\chi_A(x)\chi_B(y)\chi_B(z)]$ for general $A \subseteq B$, letting $s_a := |B \cap \pi^{-1}(a)|$,

$p_s := ((-1)^s + (1-2\delta)^s)/2$, $q_s := ((-1)^s - (1-2\delta)^s)/2$, it is equal to

$$\left(\prod_{a\in A} q_{s_a}\right)\cdot\left(\prod_{a\in\pi(B)\setminus A} p_{s_a}\right). \tag{16}$$

Note that $p_s^2 + q_s^2 \le 1-\delta$ for any $s \ge 1$. Then for fixed B,

$$\begin{aligned}\sum_{A\subseteq\pi(B)} \left(\mathbb{E}[\chi_A(x)\chi_B(y)\chi_B(z)]\right)^2 &\le \sum_{A\subseteq\pi(B)} \mathbb{E}[(\chi_A(x)\chi_B(y)\chi_B(z))^2]\\ &= \sum_{A\subseteq\pi(B)}\left(\prod_{a\in A} q_{s_a}^2 \prod_{a\in\pi(B)\setminus A} p_{s_a}^2\right)\\ &= \prod_{a\in\pi(B)} (p_{s_a}^2 + q_{s_a}^2)\\ &\le (1-\delta)^{|\pi(B)|}.\end{aligned}$$

Finally, we analyze (15). When $B=\emptyset$, we get $\widehat{f}(\emptyset)\widehat{g}(\emptyset)^2 = \alpha_i\alpha_j^2$. Say B *big* if $|B| > \delta^{-2}$ and small if $1 \le |B| \le \delta^{-2}$. Fix large B and $v \in V_j$, and consider a random edge $(u,v) \in E_{i,j}$. Since $\mathcal{M}$ is η-smooth, the probability that $|\pi(B)| \ge \delta^{-2}$ is at least $1-\eta\delta^{-4}$, so using $(1-\delta)^{1/2\delta^2} \le \delta$,

$$\mathbb{E}_{(u,v)\in E_{i,j}}(1-\delta)^{|\pi(\beta)|/2} \le \delta + \eta\delta^{-4}.$$

Therefore, for any fixed $v \in V_j$, we can bound (15) for big B as:

$$\begin{aligned}&\mathbb{E}_{(u,v)\in E_{i,j}}\left[\left|\sum_{B \text{ big}} \widehat{g}(B)^2 \sum_{A\subseteq\pi(B)} \widehat{f}(A)\mathbb{E}[\chi_A(x)\chi_B(y)\chi_B(z)]\right|\right]\\ \le\ &\mathbb{E}_{(u,v)\in E_{i,j}}\left[\sum_{B \text{ big}} \widehat{g}(B)^2\left(\sum_{A\subseteq\pi(B)}\widehat{f}(A)^2\right)^{1/2}\left(\sum_{A\subseteq\pi(B)}\mathbb{E}[\chi_A(x)\chi_B(y)\chi_B(z)]^2\right)^{1/2}\right]\\ \le\ &\mathbb{E}_{(u,v)\in E_{i,j}}\left[\sum_{B \text{ big}} \widehat{g}(B)^2(1-\delta)^{|\pi(\beta)|/2}\right] \le \delta + \eta\delta^{-4}.\end{aligned}$$

Similarly, we can bound (15) for small B and $A=\emptyset$. With probability at least $1-\eta|\pi(B)|^2 \ge 1-\eta\delta^{-4}$ we have $|\pi(\beta)| = |\beta|$, and if this happens, $\mathbb{E}[|\chi_B(y)\chi_B(z)|] \le |p_1| = \delta$. Therefore,

$$\begin{aligned}&\mathbb{E}_{(u,v)\in E_{i,j}}\left[\left|\sum_{B \text{ small}} \widehat{g}(B)^2\widehat{f}(\emptyset)\mathbb{E}[\chi_B(y)\chi_B(z)]\right|\right]\\ &\mathbb{E}_{(u,v)\in E_{i,j}}\left[\left|\sum_{B \text{ small}} \mathbb{E}[\chi_B(y)\chi_B(z)]\right|\right]\\ &\eta\delta^{-4} + \delta.\end{aligned}$$

Finally, for small B and $\emptyset \subsetneq A \subseteq \pi(B)$, we bound (15) as

$$\left|\sum_{B \text{ small}} \widehat{g}(B)^2 \sum_{\emptyset\subsetneq A\subseteq\pi(B)} \widehat{f}(A)\mathbb{E}[\chi_A(x)\chi_B(y)\chi_B(z)]\right|$$

$$\leq \left(\sum_{B \text{ small}} \widehat{g}(B)^2 \sum_{\emptyset \subsetneq A \subseteq \pi(B)} \widehat{f}(A)^2\right)^{1/2} \left(\sum_{B \text{ small}} \widehat{g}(B)^2 \sum_{\emptyset \subsetneq A \subseteq \pi(B)} \mathbb{E}[\chi_A(x)\chi_B(y)\chi_B(z)]^2\right)^{1/2}$$
$$\leq \left(\sum_{B \text{ small}} \widehat{g}(B)^2 \sum_{\emptyset \subsetneq A \subseteq \pi(B)} \widehat{f}(A)^2\right)^{1/2}.$$

Since at least an ε fraction of $e \in E_{i,j}$ is good, at least an $\varepsilon/2$ fraction of $v \in V_j$ satisfies that at least an $\varepsilon/2$ fraction of $(u, v) \in E_{i,j}$ is good. Call such v *good*. If v is good,

$$\varepsilon/2 \leq \mathbb{E}_{(u,v)\in E_{i,j}}\left[\left|\beta_{u,v} - \alpha_u \alpha_v^2\right|\right] \leq 2\delta + 2\eta\delta^{-4} + \mathbb{E}_{(u,v)\in E_{i,j}}\left[\left(\sum_{B \text{ small}} \widehat{g}(B)^2 \sum_{\emptyset \subsetneq A \subseteq \pi(B)} \widehat{f}(A)^2\right)^{1/2}\right].$$

Consider the randomized assignment where all $v \in V_j$ first chooses a set $B \subseteq R$ with probability $\widehat{g}(B)^2$ and gets random $b \in B$. (Similarly, $u \in V_i$ chooses a set $A \subseteq L$ with probability $\widehat{f}(A)^2$ and gets random $a \in A$.) Since the sum of squared Fourier coefficients is at most 1 for every f and g, this is a well-defined strategy. For good v, it will satisfy at least a

$$\frac{(\varepsilon/2 - 2\delta - 2\eta\delta^{-4})^2}{|A||B|} \leq \delta^4(\varepsilon/2 - 2\delta - 2\eta\delta^{-4})^2$$

fraction of constraints incident on v in expectation. Therefore, there is an assignment between that satisfies at least an $(\varepsilon\delta^4/2) \cdot (\varepsilon/2 - 2\delta - 2\eta\delta^{-4})^2$ fraction of edges in $E_{i,j}$. □

4.5 Make instances dense

In this section, we show how to convert hard instances to ensure $|H| = \omega(|V|^2)$ while preserving hardness, finishing the proof of Theorem 4.1. From the previous discussion, given an edge-weighted 3-hypergraph $\mathcal{H} = (V, H)$ with $n = |V|$, $m = |H|$, and $k = n/2$ (without loss of generality, assume that the sum of weights is 1), it is NP-hard to distinguish whether (1) there exists $S \subseteq V$ that intersects every hyperedge or (2) every $S \subseteq V$ with $|S| \leq n/2$ intersects hyperedges of weight at most $(7/8 + \varepsilon)$, for any constant $\varepsilon > 0$.

Let $b = \max(n, m)^\beta, c = b^{2.5}$ where β is a constant chosen later. Our reduction creates a new hypergraph $\mathcal{H}' = (V', H')$ where

- $V' = V \times [b]$.
- For each hyperedge $(u, v, w) \in H$ with weight $w(u, v, w)$, for $\lfloor c \cdot w(u, v, w) \rfloor$ times independently,
  - Sample $x, y, z \in [b]$ independently.
  - Add hyperedge $((u, x), (v, y), (w, z))$ to H'.
- If any hyperedge is added more than once, delete all occurrences of the hyperedge.

By the last step, $\mathcal{H}'$ is simple. The number of hyperedges added before the last step is at least

$$\sum_{e\in H} \lfloor c \cdot w(e) \rfloor \geq \left(\sum_{e\in H} c \cdot w(e)\right) - m = c - m.$$

Fix $(u,v,w) \in H$ and let (x,y,z) be the triple sampled in the ith iteration for (u,v,w) such that $((u,x),(v,y),(w,z))$ is added to H'. The probability that the same triple (x,y,z) is chosen in another iteration so that $((u,x),(v,y),(w,z))$ is deleted in the last step is at most c/b^3. Therefore, the expected number of hyperedges deleted in the last step of the reduction is at most $c \cdot (c/b^3)$, and the with probability at least 0.9, it is at most $c \cdot (10c/b^3)$.

Completeness. If $S \subseteq V$ intersects every hyperedge in H with nonzero weight, $S' = S \times [b]$ does the same for H'.

Soundness. For soundness, we upper bound the number of hyperedges before the last deletion step intersected by S' with $|S'| \leq |V'|/2$, because this is only an overestimate. Fix $e = (u,v,w) \in E$ and consider the hyperedges created when considering e and let $c_e = \lfloor c \cdot w(e) \rfloor$ be the number of them. Fix $X, Y, Z \in [b]$ and $\alpha_u = |X|/b$, $\alpha_v = |Y|/b$, $\alpha_w = |Z|/b$. The expected number of hyperedges not intersecting $(\{x\} \times X) \cup (\{y\} \times Y) \cup (\{z\} \times Z)$ is exactly $c_e(1-\alpha_u)(1-\alpha_v)(1-\alpha_w)$. By Hoeffding's bound, the probability that it is less than the expected value minus t is at most $\exp(-2t^2/c_e)$. By union bound, the probability that this happens for any choice of u, v, w, X, Y, Z is at most

$$m \cdot 2^{3b} \cdot \exp(-2t^2/c_e) \leq m \cdot \exp(3b - 2t^2/c).$$

By taking $t = b^{1.8}$ ensures $2t^2/c = 2b^{1.1}$ so that the above probability is at most $m\exp(-b^{1.1})$. Since b will be greater than m, this probability is $o(1)$. Therefore, with probability $1 - o(1)$, for any $(u,v,w) \in H$ and $X, Y, Z \subseteq [b]$, the number of hyperedges created from (u,v,w) not intersecting $(\{x\} \times X) \cup (\{y\} \times Y) \cup (\{z\} \times Z)$ is at least

$$c_e(1-\alpha_u)(1-\alpha_v)(1-\alpha_w) - t.$$

Then for any $S' \subseteq V$ with $|S'| = |V'|/2$, let $\alpha_v := |(\{v\} \times [b]) \cap S'|/b$ for $v \in V$, so that $\mathbb{E}_v[\alpha_v] = 1/2$. Then the total number of edges not intersecting S' is at least

$$\begin{aligned}
&\left(\sum_{e=(u,v,w) \in H} c_e(1-\alpha_u)(1-\alpha_v)(1-\alpha_w) \right) - mt \\
&\geq c \cdot \left(\sum_{e=(u,v,w) \in H} w_e(1-\alpha_u)(1-\alpha_v)(1-\alpha_w) \right) - m(t+1).
\end{aligned}$$

Note that the first term of the RHS is exactly c times the expected weight of edges of H not intersecting S, where $S \subseteq V$ is a random subset that includes $v \in V$ with probability α_v independently. By invoking the soundness condition for $\mathcal{H} = (V, H)$, the RHS is at least $c(1/8 - \varepsilon) - m(t+1)$.

Finishing up. Therefore, with probability at least $0.9 - o(1)$, $\mathcal{H}'$ has at least $c - m - 10c^2/b^3$ hyperedges. In the completeness case, there exists S with $|S| \leq |V'|/2$ that intersects every hyperedge, and in the soundness case, every S with $|S| \leq |V'|/2$ does not intersect at least $c(1/8 - \varepsilon) - m(t+1)$ hyperedges. Recall the parameter setting $c = b^{2.5}$, $t = b^{1.8}$ and $b = \max(n,m)^\beta$. Setting $\beta \geq 2$ will ensure $m(t+1) = o(c)$, so that $\mathcal{H}'$ has $(1 - o(1))c$ hyperedges and the gap is preserved to be $7/8 + \varepsilon + o(1)$. The number of vertices $|V'| = nb \leq b^{1+1/\beta}$ and the number of hyperedges $|H'| \geq (1 - o(1))c \geq \Omega(b^{2.5})$, so setting $\beta = 5$ will ensure that $|H'| = \omega(|V'|^2)$.

5 Open Problems

In this section, we list some open problems related to hardness of approximation of clustering objectives. In this regard, the most important and also immediate question is whether JCH is true?

Open Problem 5.1. *Is the Johnson Coverage Hypothesis true?*

Another important question is whether there is a 'black-box' way to ensure that the hard instance of JCH has large number of clusters. We point the reader to Section 4.5 which seems to be a first step in this direction.

Open Problem 5.2. *Does* JCH *imply* JCH**?*

Next, we move to discuss open questions with a more combinatorial-geometric flavor. We showed in Lemma 3.8 that $g_2(J(q, t+1, t)) \geq \sqrt{1 + \frac{1}{\sqrt{t^2+t}-t}}$. As t increases, the value of $\sqrt{1 + \frac{1}{\sqrt{t^2+t}-t}}$ converges to $\sqrt{3}$. However, the naive upper bound from triangle inequality (Proposition 3.5) states that $\gamma_2 \leq 3$. For small values of t we can indeed obtain an improved value: it is easy to see that $g_2(J(3, 2, 1)) = 2$ by placing the six points on the vertices of a regular hexagon in the plane. On the other hand, we suspect that $\gamma_2 \leq 2$, and confirming such a claim would also be interesting.

Open Problem 5.3. *Is* $\gamma_2 \geq 2$*?*

We provided a few lower bounds on γ_p in Section 3.2, but we are still very far from having good bounds on it.

Open Problem 5.4. *Is there a closed form expression for* γ_p *(in terms of p)?*

We now shift our attention to understanding various clustering objectives. First, we highlight that there is no inapproximability results for k-median and k-means in ℓ_∞-metric in the continuous case in $O(\log n)$-dimensions. The main obstacle that we are not able to overcome is that many natural embedding techniques create fake centers in the ℓ_∞-metric. We emphasize the requirement of $O(\log n)$-dimensions because in higher dimensions (i.e., $\text{poly}(n)$ dimensions), we were recently able to prove very strong inapproximability for k-means and k-median without candidate centers in ℓ_∞-metric [CKL21].

Open Problem 5.5. *What is the hardness of approximation for* k-means *and* k-median *in the* ℓ_∞*-metric in* $O(\log n)$ *dimensions?*

In [CKL21] we highlighted that there are inherent differences between the continuous and discrete cases for clustering problems and maybe basing hardness of clustering problems in the continuous case on JCH might not lead to a tight understanding. Elaborating, by starting from coloring problems (instead of covering problems), we obtained strong inapproximability results for clustering problems in the continuous case in the ℓ_∞-metric. It is natural to ask if this approach can be extended to other ℓ_p-metrics.

Open Problem 5.6. *Can we show inapproximability of* k-means *in the continuous case to a factor more than* $1 + 1/e$*?*

Apart from k-means and k-median, another clustering objective of interest is k-minsum (see Section 2 for the definition). However, for k-minsum in ℓ_p-metrics (for finite p), only APX-Hardness is known, leaving it open to prove stronger inapproximability results.

Open Problem 5.7. *Does* JCH *(or* JCH**) imply any non-trivial inapproximability results for* k-minsum *in* ℓ_p*-metrics?*

Finally, we discuss a more technical challenge. Our analysis of continuous case of k-median in the Euclidean metric in Theorem 3.24 is not tight and we raise the following question.

Open Problem 5.8. *Assuming* JCH**, can we show that* k-median *in Euclidean metric is hard to approximate to a factor less than* $1 + \frac{\sqrt{2}-1}{e}$*?*

Acknowledgements

We are truly grateful to Pasin Manurangsi for various detailed discussions that inspired many of the results in this paper.

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. Karthik C. S. was supported by Irit Dinur's ERC-CoG grant 772839, the Israel Science Foundation (grant number 552/16), the Len Blavatnik, the Blavatnik Family foundation, Subhash Khot's Simons Investigator Award, and by a grant from the Simons Foundation, Grant Number 825876, Awardee Thu D. Nguyen. Euiwoong Lee was supported in part by the Simons Collaboration on Algorithms and Geometry.

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.

[AJT19] Vedat Levi Alev, Fernando Granha Jeronimo, and Madhur Tulsiani. Approximating constraint satisfaction problems on high-dimensional expanders. In *FOCS*, 2019.

[AKK+08] Sanjeev Arora, Subhash Khot, Alexandra Kolla, David Steurer, Madhur Tulsiani, and Nisheeth K. Vishnoi. Unique games on expanding constraint graphs are easy: extended abstract. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing, Victoria, British Columbia, Canada, May 17-20, 2008*, pages 21–28, 2008.

[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.

[ANSW20] Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, and Justin Ward. Better guarantees for k-means and euclidean k-median by primal-dual algorithms. *SIAM J. Comput.*, 49(4), 2020.

[ARR98] Sanjeev Arora, Prabhakar Raghavan, and Satish Rao. Approximation schemes for euclidean k-medians and related problems. In *STOC*, volume 98, pages 106–113, 1998.

[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.

[BG95] Hervé Brönnimann and Michael T Goodrich. Almost optimal set covers in finite vc-dimension. *Discrete & Computational Geometry*, 14(4):463–479, 1995.

[BGJ21] Anup Bhattacharya, Dishant Goyal, and Ragesh Jaiswal. Hardness of approximation for euclidean k-median. In Mary Wootters and Laura Sanità, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2021, August 16-18, 2021, University of Washington, Seattle, Washington, USA (Virtual Conference)*, volume 207 of *LIPIcs*, pages 4:1–4:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.

[BHPI02] Mihai Bādoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 250–257. ACM, 2002.

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

[BKL12] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Hooyeon Lee. Approximating low-dimensional coverage problems. In *Proceedings of the twenty-eighth annual symposium on Computational geometry*, pages 161–170. ACM, 2012.

[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.

[CE16] Julia Chuzhoy and Alina Ene. On approximating maximum independent set of rectangles. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 820–829. IEEE, 2016.

[CGK+19] Vincent Cohen-Addad, Anupam Gupta, Amit Kumar, Euiwoong Lee, and Jason Li. Tight FPT approximations for k-median and k-means. In *46th International*

Colloquium on Automata, Languages, and Programming, ICALP 2019, July 9-12, 2019, Patras, Greece., pages 42:1–42:14, 2019.

[CGM09] Moses Charikar, Venkatesan Guruswami, and Rajsekar Manokaran. Every permutation CSP of arity 3 is approximation resistant. In *Proceedings of the 24th Annual IEEE Conference on Computational Complexity, CCC 2009, Paris, France, 15-18 July 2009*, pages 62–73, 2009.

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

[CKK+06] Shuchi Chawla, Robert Krauthgamer, Ravi Kumar, Yuval Rabani, and D. Sivakumar. On the hardness of approximating multicut and sparsest-cut. *Comput. Complex.*, 15(2):94–114, 2006.

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

[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.

[CSF19] Vincent Cohen-Addad, David Saulpic, and Andreas Emil Feldmann. Near-linear time approximations schemes for clustering in doubling metrics. In *60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019, Baltimore, Maryland, USA, November 9-12, 2019*, pages 540–559, 2019.

[DC83] Dominique De Caen. *Extension of a theorem of Moon and Moser on complete subgraphs*. Faculty of Mathematics, University of Waterloo, 1983.

[DC91] Dominique De Caen. The current status of turán's problem on hypergraphs. *Extremal problems for finite sets*, 3:187–197, 1991.

[DF95] Rodney G. Downey and Michael R. Fellows. Fixed-parameter tractability and completeness II: on completeness for W[1]. *Theor. Comput. Sci.*, 141(1&2):109–131, 1995.

[DGKR05] Irit Dinur, Venkatesan Guruswami, Subhash Khot, and Oded Regev. A new multilayered PCP and the hardness of hypergraph vertex cover. *SIAM J. Comput.*, 34(5):1129–1146, 2005.

[Din07] Irit Dinur. The PCP theorem by gap amplification. *J. ACM*, 54(3):12, 2007.

[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.

[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.

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

[FL11] Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 569–578. ACM, 2011.

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

[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.

[GL17] Badih Ghazi and Euiwoong Lee. Lp/sdp hierarchy lower bounds for decoding random ldpc codes. *IEEE Transactions on Information Theory*, 64(6):4423–4437, 2017.

[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.

[GS20] Venkatesan Guruswami and Sai Sandeep. An algorithmic study of the hypergraph turán problem. *CoRR*, abs/2008.07344, 2020.

[Hås01] Johan Håstad. Some optimal inapproximability results. *J. ACM*, 48(4):798–859, 2001.

[HS93] D. A. Holton and J. Sheehan. *The Johnson graphs and even graphs*, page 300. Australian Mathematical Society Lecture Series, 7, Cambridge: Cambridge University Press, 1993.

[IP01] Russell Impagliazzo and Ramamohan Paturi. On the complexity of k-SAT. *J. Comput. Syst. Sci.*, 62(2):367–375, 2001.

[IPZ01] Russell Impagliazzo, Ramamohan Paturi, and Francis Zane. Which problems have strongly exponential complexity? *J. Comput. Syst. Sci.*, 63(4):512–530, 2001.

[JMS02] Kamal Jain, Mohammad Mahdian, and Amin Saberi. A new greedy approach for facility location problems. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 731–740, 2002.

[Kee11] Peter Keevash. Hypergraph turan problems. *Surveys in combinatorics*, 392:83–140, 2011.

[Kho02] Subhash Khot. Hardness results for coloring 3 -colorable 3 -uniform hypergraphs. In *43rd Symposium on Foundations of Computer Science (FOCS 2002), 16-19 November 2002, Vancouver, BC, Canada, Proceedings*, pages 23–32. IEEE Computer Society, 2002.

[KKMO07] Subhash Khot, Guy Kindler, Elchanan Mossel, and Ryan O'Donnell. Optimal inapproximability results for MAX-CUT and other 2-variable csps? *SIAM J. Comput.*, 37(1):319–357, 2007.

[KLM19] Karthik C. S., Bundit Laekhanukit, and Pasin Manurangsi. On the parameterized complexity of approximating dominating set. *J. ACM*, 66(5):33:1–33:38, 2019.

[KM20] Karthik C. S. and Pasin Manurangsi. On closest pair in euclidean metric: Monochromatic is as hard as bichromatic. *Comb.*, 40(4):539–573, 2020.

[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.

[KR08] Subhash Khot and Oded Regev. Vertex cover might be hard to approximate to within 2-epsilon. *J. Comput. Syst. Sci.*, 74(3):335–349, 2008.

[KSS10] Amit Kumar, Yogish Sabharwal, and Sandeep Sen. Linear-time approximation schemes for clustering problems in any dimensions. *Journal of the ACM (JACM)*, 57(2):5, 2010.

[KV15] Subhash Khot and Nisheeth K. Vishnoi. The unique games conjecture, integrality gap for cut problems and embeddability of negative-type metrics into l_1. *J. ACM*, 62(1):8:1–8:39, 2015.

[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.

[Man20] Pasin Manurangsi. Tight running time lower bounds for strong inapproximability of maximum k-coverage, unique set cover and related problems (via t-wise agreement testing theorem). In *SODA*, 2020.

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

[Nik11] Vladimir Nikiforov. The number of cliques in graphs of given order and size. *Transactions of the American Mathematical Society*, 363(3):1599–1618, 2011.

[O'D14] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

[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.

[Raz08] Alexander A Razborov. On the minimal density of triangles in graphs. *Combinatorics, Probability and Computing*, 17(4):603–618, 2008.

[Rei16] Christian Reiher. The clique density theorem. *Annals of Mathematics*, pages 683–707, 2016.

[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.

[Sid95] Alexander Sidorenko. What we know and what we do not know about turán numbers. *Graphs and Combinatorics*, 11(2):179–199, 1995.

[Sid97] Alexander Sidorenko. Upper bounds for turán numbers. *journal of combinatorial theory, Series A*, 77(1):134–147, 1997.

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

[Tur41] Pal Turan. Research problems. *Maguar Tud. Akad. Mat. Kutato Int. Közl.*, 6:417–423, 1941.

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

A Obstacles that need to be Overcome to prove Johnson Coverage Hypothesis

In this section, we consider a few natural approaches to proving JCH, and demonstrate central obstacles to proceed using this approach.

It is well known that Max k-Coverage is hard to approximate to factor beyond $1 - 1/e$ and one may wonder if there are any "simple" gap preserving reductions from Max k-Coverage to Johnson Coverage problem. In the below theorem, we show that if there is such a reduction then there should be a significant blowup in size of the witness. In fact our result is stronger, as we show that even to reduce gap Max k-Coverage to exact Johnson Coverage problem (i.e., at the cost of losing the all the gap in Max k-Coverage), we still need to blow up in witness size.

Theorem A.1. *Let $\alpha, \delta \geq 0$. Let z be some fixed constant. Suppose there is an algorithm $\mathcal{A}$ that on given as input a Max k-Coverage instance $(\mathcal{U}, \mathcal{S}, k)$, outputs a (α, z)-Johnson Coverage instance $([n], E, k')$ such that the following holds.*

Size: *$|\mathcal{U}| = |\mathcal{S}|^{O(1)} = n^{O(1)}$ and $k' = F(k)$ for some computable function F.*

Completeness: *If there are k sets in $\mathcal{S}$ that cover $\mathcal{U}$ then there exists $\mathcal{C} := \{S_1, \ldots, S_{k'}\} \subseteq \binom{[n]}{z-1}$ such that*

$$\mathsf{cov}(\mathcal{C}) := \bigcup_{i \in [k']} \mathsf{cov}(S_i, E) = E.$$

Soundness: *If no k sets in $\mathcal{S}$ cover $(1 - \delta)$ fraction of $\mathcal{U}$ then for every $\mathcal{C} := \{S_1, \ldots, S_{k'}\} \subseteq \binom{[n]}{z-1}$ we have $|\mathsf{cov}(\mathcal{C})| \leq \alpha \cdot |E|$.*

Running Time: *$\mathcal{A}$ runs in time $T(k) \cdot poly(n)$ for computable function T.*

Then the following consequences hold:

- *If $\delta = 0$ then W[2]=FPT.*
- *If $\delta \leq 1/k$ then W[1]=FPT.*
- *If $\delta \leq 1/e - \varepsilon$ for some constant $\varepsilon > 0$ then* Gap-ETH *is false.*

In order to understand the consequences, we remark the following:

Remark A.2. *Note that $W[2] \neq FPT$ is a weaker assumption than $W[1] \neq FPT$, which is in turn a weaker assumption than* ETH *and consequently* Gap-ETH.

The proof of Theorem A.1 follows from Lemma A.3, Theorem A.4, Theorem A.6, and a result of [DF95].

Lemma A.3. *There is an algorithm running in time $z^k \cdot n^{O(1)}$ that can decide any $(0, z)$-Johnson Coverage instance.*

Proof. We essentially follow the same idea as the FPT algorithm for vertex cover. Let $([n], E, k)$ be a $(0, z)$-Johnson Coverage instance. Pick an arbitrary set $S \in E$. There are z possible subsets in $\binom{[n]}{z-1}$, say $T_1, \ldots T_z$, that can cover S. We branch and consider all z possibilities. Suppose we branch and pick T_i. Then we remove all subsets in E that are covered by T_i, and repeat by

picking another arbitrary subset S' in E. We stop the algorithm after the branching tree is of height k. If there are k subsets in $\binom{[n]}{z-1}$ that cover all subsets in E, then in one of the branching, we would have found it. The size of the branching tree is at most z^k. and at each step, updating E can be done in linear time. □

Theorem A.4 (Essentially [KLM19]). *Given an instance $(\mathcal{U}, \mathcal{S}, k)$ of Max k-Coverage, it is W[1]-Hard (when parameterized by k) to distinguish between the following two cases:*

Completeness: *There are k sets in $\mathcal{S}$ that cover $\mathcal{U}$.*

Soundness: *No k sets in $\mathcal{S}$ cover $\left(1 - \frac{1}{k}\right)$ fraction of $\mathcal{U}$.*

Proof Sketch. We use the notation and terminology of MaxCover problem given in [KLM19]. Our starting point is the following result.

Theorem A.5 ([KLM19]). *For every $\varepsilon > 0$, given an instance Γ of MaxCover as input, it is $W[1]$-Hard to distinguish between the following two cases:*

Completeness: *MaxCover(Γ) = 1.*

Soundness: *MaxCover(Γ) $\leq \varepsilon$.*

Moreover, this holds even when the size of each left super node is a constant only depending on k and ε.

Then we can simply apply Feige's [Fei98] proof framework to conclude the corollary; details follow. Let the input instance to Maxcover be $\Gamma := (U := U_1 \cup \cdots \cup U_q, W := W_1 \cup \cdots \cup W_k, E)$. We build an instance $(\mathcal{U}, \mathcal{S}, k)$ of Max k-Coverage as follows:

$$\mathcal{U} := \{(i, f) \mid i \in [t], f : U_i \to [k]\}, \mathcal{S} := \{S_{j,w} \mid j \in [k], w \in W_j\}, \text{ and}$$

$$(i, f) \in S_{j,w} \Leftrightarrow \exists u \in U_i \text{ such that } f(u) = j \text{ and } (u, w) \in E.$$

Its easy to see that a labeling of W corresponds to a k-tuple of sets in $\mathcal{S}$. In the completeness case, a labeling of W which yields $\text{MaxCover}(\Gamma) = 1$ also corresponds to k sets in $\mathcal{S}$ that cover all of $\mathcal{U}$. In the soundness case, we claim that there are at least $(1 - \varepsilon) \cdot q$ many universe elements in $\mathcal{U}$ that are not covered by any k sets in $\mathcal{S}$. Note that $|\mathcal{U}| = \sum_{i \in [q]} k^{|U_i|} = O_{k,\varepsilon}(q)$. With the right choice of ε, the theorem statement follows. □

Theorem A.6 ([CGK+19, Man20]). *Assuming* Gap-ETH, *for every $\varepsilon > 0$, there is no algorithm running in time $n^{o(k)}$ which given an instance $([n], \mathcal{S}, k)$ of Max k-Coverage, can distinguish between the following two cases:*

Completeness: *There are k sets in $\mathcal{S}$ that cover $[n]$.*

Soundness: *No k sets in $\mathcal{S}$ cover $\left(1 - \frac{1}{e} + \varepsilon\right)$ fraction of $[n]$.*

Proof of Theorem A.1. Suppose there is an algorithm $\mathcal{A}$ as claimed in the theorem statement. Let $(\mathcal{U}, \mathcal{S}, k)$ be a Max k-Coverage instance. We run $\mathcal{A}$ to obtain a (α, z)-Johnson Coverage instance $([n], E, k')$. We then run the algorithm in Lemma A.3 to distinguish if $([n], E, k')$ is part of the completeness case or the soundness case. Thus we obtained an algorithm for deciding $(\mathcal{U}, \mathcal{S}, k)$ that runs in time $T'(k) \cdot \text{poly}(|\mathcal{U}|)$ for some computable function T'. Depending on the value of δ, this contradicts either W[2]-Hardness of Max k-Coverage [DF95], Theorem A.4, or Theorem A.6. □

It's worth noting that Theorem A.1 only holds for constant z. In fact, if it suffices to only show Gap-ETH is false as a consequence in Theorem A.1 then we can allow z to be $n^{o(1)}$. On the other hand, we could completely get away with any kind of restriction on z by setting $\delta > 0$ in Theorem A.1 and invoking the result in [BKL12] instead of Lemma A.3.

Next, we show that in order to prove JCH, we would require to prove NP-hardness of a plausibly highly structured Label Cover instance. We first recaptiluate here proof of the $(1 - 1/e)$-factor inapproximability of Max k-Coverage shown by Feige [Fei98]. We present the proof outline below in terms of label cover (as in [Mos15, DS14]) instead of multi-prover proof systems (as in [LY94, Fei98]).

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.7 (Label Cover problem[8])**.** *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 $\mathsf{sat}(\Pi, \sigma) \leq \varepsilon$.*

An immediate consequence of the PCP theorem is that it is NP-hard to decide an instance $\Pi(G, \pi)$ of $(1 - \varepsilon, d, \alpha)$-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, and followed by a strengthening of the soundness guarantee due to Moshkovitz [Mos15] we get the following.

Theorem A.8 (Bounded Label Cover Inapproximability [AS98, ALM+98, Raz98, Mos15])**.** *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. Moreover, this holds even for the following soundness guarantee: for every assignment $\sigma_U : U \to \Sigma_U$ we have the following:*

$$\mathsf{weak\text{-}sat}(\sigma_U) := \frac{|\{v \in V \mid \exists u_1 \neq u_2 \in N(v), \pi_{u_1,v}(\sigma(u_1)) = \pi_{u_2,v}(\sigma(u_2))\}|}{|V|} \leq \varepsilon$$

We define a gadget relevant to Max k-Coverage:

[8]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.

Definition A.9. *Let $d, q \in \mathbb{N}$ and $S \subseteq [d]^q$. We say that S is (d,q)-resistant if the following holds: for every $H_1, \cdots, H_d$ be d axis-parallel hyperplanes such that no two are mutually parallel, there is a point in S not contained in any of the hyperplanes.*

Theorem A.10 (Essentially Feige [Fei98])**.** *Let $\Pi(G, \pi)$ be a hard instance of (ε, d, α)-label cover problem as in Theorem A.8 and let S be a $(d, |\Sigma_V|)$-resistant set. Consider the following set-system $(\mathcal{U}, \mathcal{S})$:*

$$\mathcal{U} := \{(v,s) \mid v \in V, s \in S\}, \mathcal{S} := \{S_{u,a} \mid u \in U, a \in \Sigma_U\}, \text{ and}$$

$$(v,s) \in S_{u,a} \Leftrightarrow (u,v) \in E \text{ and } s_{\pi_{(u,v)}(a)} = v.$$

Then, it is NP*-Hard to distinguish between the following:*

Completeness: *There are $|U|$ sets in $\mathcal{S}$ that cover $\mathcal{U}$.*

Soundness: *No $|U|$ sets in $\mathcal{S}$ cover $(1-\delta)$ fraction of $\mathcal{U}$ for some positive constant δ only depending on ε, $|\Sigma_U|$, $|\Sigma_V|$, and d.*

In fact by using S to be the entire space $[d]^{|\Sigma_V|}$ and with a more tighther analysis, Feige showed the tight $1 - 1/e$ inapproximability of Max k-Coverage. He even noted that a random subset S of small cardinality (roughtly $\widetilde{O}(d|\Sigma_V|)$) suffices. We show below that there is no such subset S for which we could use Feige's framework to obtain JCH.

Theorem A.11. *Let $\Pi(G, \pi)$ be a hard instance of (ε, d, α)-label cover problem as in Theorem A.8 and let S be a $(d, |\Sigma_V|)$-resistant set. Consider the following set-system $(\mathcal{U}, \mathcal{S})$:*

$$\mathcal{U} := \{(v,s) \mid v \in V, s \in S\}, \mathcal{S} := \{S_{u,a} \mid u \in U, a \in \Sigma_U\}, \text{ and}$$

$$(v,s) \in S_{u,a} \Leftrightarrow (u,v) \in E \text{ and } s_{\pi_{(u,v)}(a)} = v.$$

Suppose $(\mathcal{U}, \mathcal{S})$ is an instance of $(0, z)$-Johnson Coverage problem (for some constant z) then Gap-ETH *is false.*

The proof follows by noting the following lower bound on parameterized label cover problem and Lemma A.3.

Theorem A.12 (Parameterized Label Cover Inapproximability with total disagreement [Man20])**.** *Assuming* Gap-ETH*, for every constant $\varepsilon > 0$, there exist constants $d := d(\varepsilon) \in \mathbb{N}$ and $\alpha := \alpha(\varepsilon) \in \mathbb{N}$ such that no algorithm running in time $\alpha^{o(k)}$ given as input a instance $\Pi(G, \pi)$ of (ε, d, α)-label cover problem (where $|U| = k$), can 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 $\sigma_U : U \to \Sigma_U$ we have* $\mathsf{weak\text{-}sat}(\sigma_U) \leq \varepsilon$.

ECCC ISSN 1433-8092
https://eccc.weizmann.ac.il