Electronic Colloquium on Computational Complexity, Revision 2 of Report No. 31 (2013)

Clustering in the Boolean Hypercube in a List Decoding Regime

Irit Dinur * Elazar Goldenberg †

April 29, 2013

Abstract

We consider the following *clustering with outliers* problem: Given a set of points $X \subset \{-1,1\}^n$, such that there is some point $z \in \{-1,1\}^n$ for which $\Pr_{x \in X}[\langle x, z\rangle \geq \varepsilon] \geq \delta$, find z. We call such a point z a (δ, ε)-center of X.

In this work we give lower and upper bounds for the task of finding a (δ, ε)-center. We first show that for $\delta = 1-\nu$ close to 1, i.e. in the "unique decoding regime", given a $(1-\nu, \varepsilon)$-centered set our algorithm can find a $(1-(1+o(1))\nu, (1-o(1))\varepsilon)$-center. More interestingly, we study the "list decoding regime", i.e. when δ is close to 0. Our main upper bound shows that for values of ε and δ that are larger than $1/\mathrm{poly}\log(n)$, there exists a polynomial time algorithm that finds a $(\delta - o(1), \varepsilon - o(1))$-center. Moreover, our algorithm outputs a list of centers explaining all of the clusters in the input.

Our main lower bound shows that given a set for which there exists a (δ, ε)-center, it is hard to find even a $(\delta/n^c, \varepsilon)$-center for some constant c and $\varepsilon = 1/\mathrm{poly}(n), \delta = 1/\mathrm{poly}(n)$.

*Weizmann Institute of Science and Radcliffe Institute for Advanced Study. Research supported in part by the Israel Science Foundation grant no. 1179/09 and by the Binational Science Foundation grant no. 2008293 and by an ERC grant no. 239985.

†Weizmann Institute of Science.

ISSN 1433-8092

1 Introduction

Suppose we are given access to a set of points $X \subset \{-1,1\}^n$ such that at least δ fraction of these points are ε-correlated with some unknown "center" $z \in \{-1,1\}^n$. We wish to recover (an approximation of) z even if the remaining $1-\delta$ fraction of the points in X are arranged in an adversarial manner. Formally, a (δ,ε)-center is defined as follows,

Definition 1.1. Given a set $X \subset \{-1,1\}^n$, the point $z \in \{-1,1\}^n$ is called a (δ,ε)-center if there exists $X' \subset X$, $|X| \geq \delta X$, such that:

$$\forall x \in X' \, \langle x, z \rangle \geq \varepsilon.$$

We denote by $C_\varepsilon(z)$ the set of all points $x \in X$ satisfying $\langle z, x \rangle \geq \varepsilon$.

We call $C_\varepsilon(z)$ the *cluster* of z in X.

Clustering is a vastly studied topic, but usually the the focus is on inputs that are drawn from some unknown (parameterized) distribution, or on deterministic data with a *small* amount of adversarial noise. Here we consider the problem in the "list decoding" regime, where the fraction of corrupted data points approaches 1. In this case, there are potentially more than a single cluster, so the algorithm needs to output a *list* of clusters.

Formally, we study the following (δ,ε)-clustering problem: Given a set $X \subset \{-1,1\}^n$ that contains a (δ,ε)-center, find all such centers.

Ideally, we are seeking an algorithm that lists all possible centers. Of course, list decoding is feasible when there is some way to bound the list size. In the case of error correcting codes, the distance of the code may facilitate such a bound. In our case, there is no underlying code, so we instead rely on an approximate representation. The idea is simply to output a short list of centers such that every cluster is 'represented' by some center in the list. Pinning down the best notion of 'representation' turns out to be tricky, and we view this as part of the contribution of this paper, on which we elaborate more in Section 3.2.

In this work, we study upper and lower bounds for the problem of finding (δ,ε)-center. The complexity of this problem depends on the choice of the parameters ε and δ: We show that when ε or δ are close to 1, then the task of finding a (δ,ε)-center is relatively easy. For ε and δ that are larger than $1/\operatorname{poly}\log(n)$ we present an algorithm that finds a $((1-o(1))\delta,(1-o(1))\varepsilon)$-center (**Elazar:** assuming a certain structure on the set X). We complement our results by showing hardness results when δ and ε are much smaller. We elaborate on these results next.

1.1 Upper Bounds

In this part we present several approximation results for the (δ,ε)-clustering problem. The approximation version of the (δ,ε)-center problem allows the output to be a center whose cluster has a smaller margin value $\varepsilon' \leq \varepsilon$, and that contains a smaller $\delta' \leq \delta$ fraction of the points. In other words, under the promise of existence of a (δ,ε) center, the approximation algorithm will find a (δ',ε')-center.

Definition 1.2 (Approximate-Cluster Problem)**.** An instance of the problem is a (δ,ε)-centered set of points $X = \{x_1,\ldots,x_N\} \subset \{-1,1\}^n$. The goal is to find a (δ',ε') center with parameters as close as possible to ε,δ.

We first give an approximation algorithm for the easier "unique-decoding" parameter regime, i.e. where δ is close to 1.

Theorem 1.3. *Let $0 < \varepsilon, \delta < 1$, and let $X \subset \{-1,1\}^n$. There is a polynomial-time algorithm for solving the following problems,*

1. *If X is $(1, \varepsilon)$-centered find a $(1, \varepsilon - O(\frac{\log N}{\sqrt{n}}))$-center.*
2. *If X is $(1-\nu, \varepsilon)$-centered find a $(1 - 1/a, \varepsilon - a\nu(1+\varepsilon))$-center, for any $a > 1$. In particular, for a parameter $\tau > 0$ if $\nu < \tau^2\varepsilon/(1+\varepsilon)$ then this gives a $(1-\tau, (1-\tau)\varepsilon)$-center.*

This algorithm is simply based on linear programming.

We next turn to the more challenging setting that is when both ε and δ are small. Our main result is a polynomial time algorithm that approximates the (δ, ε)-center for values of ε and δ that are larger than $1/\text{poly}\log(n)$. As explained earlier we are in a "list decoding" setting, that allows for more than one cluster to exist simultaneously. Moreover, we want our algorithm to output a list that "exhaustively" explains all of the clusters in the data.

Here, the goal for the algorithm is to output an exhaustive list of centers, namely a list for which:

- Each member in the list is a cluster in the data.
- Each cluster in the data is approximately equal to one of the clusters in the list.

The reason for asking only for an approximate equality to members in the list is clear: there can be an exponential number of different clusters, and approximation seems like a natural way to get a manageable list size. However, it turns out that even when allowing approximate centers, there still might be an exponential number of them (more details in Section 3.2). We show that this can only occur if the exponentially-many clusters are contained in one bigger cluster. In light of this example, the new goal for the algorithm becomes to output a list of centers that "cover" all of the clusters in the set. We state below an informal version of our main theorem, for a formal version please see Section 3.2.

Theorem 1.4 (Main result, informal)**.** *Let $\varepsilon, \delta > 0$ be parameters, and let $X \subset \{-1,1\}^n$ be (δ, ε)-centered, with $|X| = N$. There exists an algorithm that runs in time polynomial in $n, N, \exp(\frac{1}{\varepsilon^2 \log 1/\varepsilon\delta})$, and outputs a list L of points each in $\{-1,1\}^n$, such that with probability $1 - 2^{-n}$ the following holds:*

- *Each $y \in L$ is a $((1-o(1))\cdot\delta, (1-o(1))\cdot\varepsilon)$ center for X.*
- *For every $z \in \{-1,1\}^n$ which is a (δ, ε)-center and , there exists $y \in L$ such that,*

$$C_\varepsilon(z) \subset_{o(1)} C_{(1-o(1))\varepsilon}(y),$$

  where for sets A, B and $0 < \tau < 1$ we define $A \subseteq_\tau B$ if $|A \setminus B| < \tau |A|$.

- *Moreover, if $z \in \{-1,1\}^n$ is a (δ, ε)-center that is approximately maximal[1], there exists $y \in L$ such that,*

$$\left|C_\varepsilon(z) \triangle C_{(1-o(1))\varepsilon}(y)\right| < o(1)\left|C_\varepsilon(z)\right|.$$

We note that, while a priory the number of covering centers could be exponentially large, the correctness of our algorithm is a proof that it is polynomially bounded.

Let us briefly sketch our proof of this theorem. We first randomly restrict the given set of points into a small poly-logarithmic subset of coordinates. We show that a (δ, ε)-center for X is still a

[1] Essentially, it is not approximately contained in any larger cluster, see Definition 3.2

center in the restricted space (if δ and ε are large enough). Therefore, we can enumerate over all possible centers and find a solution in the restricted space. Then we show how to extend the solution from the local space into a global solution.

Our last algorithmic result deals with smaller values of ε, $\varepsilon \geq \log n/\sqrt{n}$. When δ and ε are this small the above algorithm runs in super polynomial time. For such a choice of parameters, we prove that there is always a *data point* $x \in X$ that is itself a $(2\delta\varepsilon, \varepsilon^2)$-center. This leads to the following algorithm:

Theorem 1.5. *Let $X \subset \{-1,1\}^n$ be an (δ,ε)-centered set of N elements and let $\varepsilon \gg \log n/\sqrt{n}$. There is an algorithm that runs in time* $\mathrm{poly}(N)$ *and outputs a list L such that:*

- *Each $z \in L$ is a $(2\delta\varepsilon, \varepsilon^2)$-center.*
- *For every center z, there exists a center $z' \in L$, such that at least 2ε fraction of the points $x \in X$ satisfying $\langle x, z\rangle > \varepsilon$ are also satisfying $\langle x, z'\rangle > \varepsilon^2$.*

Observe that this result is incomparable to the one attained in Theorem 1.4. While in Theorem 1.4 we are able to find almost the whole cluster of each maximal center, the algorithm proposed by Theorem 1.5 finds only a non-trivial subset of each cluster. On the other hand, Theorem 1.5 is stronger in the sense that it has a guarantee for each center, and not just for a subset of them as in Theorem 1.4.

1.2 Lower Bounds

We next turn to lower bounds. It is not hard to see that given a $(1, \mathrm{poly}(1/n))$-centered set it is **NP**-hard to find such a center, by reduction from, say, 3SAT. Moreover, we describe stronger reductions that show the hardness of the approximation problem. That is, we show that for some choices of the parameters $(\delta, \varepsilon, \delta', \varepsilon')$ the approximate center problem is infeasible unless $\mathcal{BPP} \supseteq \mathbf{NP}$. Formally, we consider the following gap-clustering problem:

Definition 1.6. The gap-clustering problem with parameters $(\delta, \varepsilon, \delta', \varepsilon')$:

The input of the problem is a set of point $X \subset \{-1,1\}^n$. The goal is to distinguish between the following cases:

- There exists a (δ,ε)-center in X.
- There is no (δ',ε')-center in X.

There are four parameters involved so it is complicated to understand the tradeoffs between the settings of the parameters. There are two key points to address: First we would like to get as large as possible a gap between δ and δ' and ε and ε'. The second is the location of the gap: find the largest ε and δ for which the problem is still hard.

Since a large gap between δ and δ' might lead to a small gap between ε and ε', and vice versa, we separate this optimization question into two: find largest δ-gap and find the largest ε-gap.

Our first hardness result focuses on the gap between ε and ε'. It shows that it is hard to distinguish between the case that there exists a (δ,ε)-center, and the case that there is no $(\delta/c, \varepsilon/2)$-center for some constant $c > 1$, ε which is an arbitrarily large polynomial in $1/n$, and δ that is a constant, formally:

Theorem 1.7. *Unless $\mathcal{BPP} \supseteq \mathbf{NP}$, there exist constants $\delta, c > 1$, such that for every constant $\alpha > 2$, it is infeasible to solve the gap-clustering problem with parameters: $\delta, \delta' = \delta/c$, $\varepsilon = \frac{2}{n^{1/\alpha}} - o(\frac{1}{n^{1/\alpha}})$ and $\varepsilon' = \frac{1}{n^{1/\alpha}} + \omega(\frac{1}{n^{1/\alpha}})$.*

Our next result focuses on amplifying the gap between δ and δ'. It shows that it is **NP**-hard to distinguish between the case that there exists a (δ, ε)-center, and the case that there is no $(\frac{\delta}{n^c}, \varepsilon')$-center, for some constant c, and for ε, δ which are poly$(1/n)$, and $\varepsilon' = (1 - o(1))\varepsilon$. Formally:

Theorem 1.8. *Unless $\mathcal{BPP} \supseteq$ **NP**, there exist constants $c_1 > 0, c_2 > 0$ such that it is infeasible to solve the gap-clustering problem with parameters: $\delta = n^{-c_1}, \delta' = \frac{\delta}{n^{c_2}}$ and $\varepsilon > \varepsilon' = \Theta(n^{-1/3})$.*

There is a gap between our algorithmic results and the aforementioned lower bounds, two particular open questions are:

- Theorem 1.5 states that given a (δ, ε)-centered set, there is a polynomial time algorithm that finds a $(\delta\varepsilon, \varepsilon^2)$-center. A natural question that arises is how hard is the task of finding a better center - that is finding a (δ', ε')-center for $\varepsilon' \gg \varepsilon^2$ and δ' being non trivial.
- Both our hardness results deal with sub-constant values of ε, and it is not clear whether we can strengthen our hardness result to deal with larger values of ε. In particular, given a $(\delta = 1/\text{poly}(n), \varepsilon = \Omega(1))$-centered set is it hard to find a $(\delta', \varepsilon/2)$-center for any nontrivial δ' ? Note that if we take δ to be larger than $1/\text{poly}(\log(n))$, then by Theorem 1.4 we can find an approximate solution in polynomial time.

1.3 Related Work

1.3.1 Upper Bounds

The most related work on clustering with outliers, as far as we know, is the work of [BHPI02]. This work considers several clustering problems, one of which is the clustering with outliers problem. The main difference is that we consider a set of data points in the Boolean hypercube $\{-1, 1\}^n$, whereas they consider a set of points in $\mathbb{R}^n$, and their algorithm outputs centers that are in $\mathbb{R}^n$ as well. We provide a more detailed comparison between our work and theirs in Appendix A.

We are not aware of works that looked at the "list-decoding" version of the clustering problem, where the algorithm needs to output a list explaining all of the clusters in the data. In other settings in theoretical computer science list decoding has been, of course, extremely successful. The seminal work of Goldreich and Levin [GL89] has a similar feel: a string is recovered from its noisy parities. In our setting, we get not a parity of the bits, but an ε-correlated version of them, and this only on a δ fraction of the inputs. The most similar works are those of [IJK06, IJKW10] on list decoding direct products. In these works the decoding algorithm is given an access to k-tuples of bits of the hidden string such that only a δ fraction of the k-tuples are correct and the rest are adversarial noise. Our setting is even harder in that even the δ fraction of "good" inputs are only guaranteed to be ε-correlated with the hidden string, rather than completely equal to it on k known bits. Extending our clustering algorithm to a direct product decoding result has been one of our main motivations, and is still work in progress.

1.3.2 Lower Bounds

Our lower bounds are closely related to the works of [FGKP09, GR06] on the MaxLin-$\mathbb{Q}$ problem, defined as follows: Given a system of equations over the rationales, and we are expected to "almost" satisfy as many equations as possible. Formally a MaxLin-$\mathbb{Q}$ with parameters $(N, n, \delta, \varepsilon)$ consists of a system of N equations over n variables $x_1, \ldots, x_n$ with rational coefficients,

$$\{a_{i0} + \sum_{j=1}^{N} a_{ij} x_j = 0\}_{j=1,\ldots,N}$$

and the goal is to distinguish between the following cases:

- At least $(1-\delta)N$ of the equations can be satisfied.
- In any assignment:
$$\left| a_{i0} + \sum_{j=1}^{N} a_{ij} x_j \right| < \varepsilon$$
is true for at most δN equations.

In [FGKP09] this problem is shown to be **NP**-hard for any constant value of $\delta > 0$.

The gap-clustering problem and MaxLin-$\mathbb{Q}$ are similar in the following sense: In the completeness case, there exits an assignment (center) that satisfies (correlates) much more equations (points) compared to the soundness case. Furthermore, the quality of the solution considered in the completeness is much better compared to the soundness case. However, there are several hurdles that prevent us from reducing MaxLin-$\mathbb{Q}$ into gap-clustering. First, the coefficients of the linear-equations can take values outside $\{-1, 1\}$ unlike in gap-clustering. Second, in MaxLin-$\mathbb{Q}$ we are trying to satisfy equalities, and not inequalities as in gap-clustering. Third, note that it is hard to solve MaxLin-$\mathbb{Q}$ even when there is an assignment that satisfies $1-\varepsilon$ fraction of equalities. In comparison, the problem of finding a $(1-\delta, \varepsilon)$-center is easy, see Theorem 1.3.

Although we could not directly reduce MaxLin-$\mathbb{Q}$ into gap-clustering, we were able to apply similar ideas to those presented in [FGKP09] and [GR06] to derive our hardness results.

Organization of the paper: Section 2 contains standard tools we use later. Section 3 studies the upper bounds for our clustering problem and contains the proofs of Theorem 1.3, Theorem 1.4, and Theorem 1.5. In Section 4 we study lower bounds for the gap-clustering problem and prove Theorem 1.7 and Theorem 1.8. We conclude by Section 5 showing information theoretic bounds on the list size of all (δ, ε)-centers.

2 Preliminaries

We state the Johnson bound as appears in the book [AB09]. It asserts that, for an error correcting code with distance $1/2 - \varepsilon^2$, and for every word x, a ball of radius $1/2 - \varepsilon$ around x cannot contain too many codewords.

Lemma 2.1 (Johnson Bound [Joh62], Theorem 19.23 in [AB09]). *Let $0 < \varepsilon < 1$, for every $x \in \{0,1\}^n$, there exist at most $1/(2\varepsilon)$ vectors $y_1, \ldots, y_\ell \in \{0,1\}^n$ such that $\Delta(x, y_i) \leq 1/2 - \varepsilon$ for every $i \in [\ell]$, and $\Delta(y_i, y_{i'}) \geq 1/2 - \varepsilon^2$ for every $i \neq i' \in [\ell]$.*

We also state here the standard Chernoff bound:

Lemma 2.2 (Chernoff Bound). *Let $X_1, \ldots, X_t$ be random independent variables taking values in the interval $[0,1]$, with expectations $\mu_1, \ldots, \mu_t$, respectively. Let $X = \frac{1}{t}\sum_{i\in[t]} X_i$, and let $\mu = \frac{1}{t}\sum_{i\in[t]} \mu_i$ be the expectation of X. For any $0 < \gamma \leq 1$, we have the following:*

$$Pr[|X - \mu| \geq \gamma] \leq \exp^{-\gamma^2 n/3}.$$

Notation. For two sets $A, B \subseteq \{-1,1\}^n$ we denote their symmetric difference by $A \triangle B$. For a vector $z \in \{-1,1\}^n$ and a subset $K \subseteq [n]$, we denote by z_K its restriction to the coordinates in K.

3 Upper Bounds: algorithms for clustering

In this section we describe algorithms for clustering first in the "unique decoding" regime, where a small fraction of the data points are corrupted, and then in the "list decoding" regime, where a very large fraction of the data is corrupted.

We first observe that if X has a (δ, ε)-center, for $\varepsilon = 1-\tau$ for small τ, then finding an approximate center is relatively easy: Any point $x \in X$ that belongs to the centered cluster is itself a $(\delta, 1-2\tau)$-center for that cluster, by the triangle inequality. By enumerating over all elements in X and checking for each $x \in X$ how many $y \in X$ are within the specified radius, we can recover a $(\delta, 1-2\tau)$-center.

The more interesting case is, therefore, when $\varepsilon < 1/2$ approaches 0.

3.1 Clustering with Few Outliers (Proof of Theorem 1.3)

We begin by addressing the easier "unique decoding" regime, where only a relatively small fraction of the data points are corrupt. More accurately, we give an algorithm that addresses the situation where δ is close to 1.

Theorem 1.3. *Let $0 < \varepsilon, \delta < 1$, and let $X \subset \{-1,1\}^n$. There is a polynomial-time algorithm for solving the following problems,*

1. *If X is $(1, \varepsilon)$-centered find a $(1, \varepsilon - O(\frac{\log N}{\sqrt{n}}))$-center.*
2. *If X is $(1-\nu, \varepsilon)$-centered find a $(1 - 1/a, \varepsilon - a\nu(1+\varepsilon))$-center, for any $a > 1$. In particular, for a parameter $\tau > 0$ if $\nu < \tau^2\varepsilon/(1+\varepsilon)$ then this gives a $(1-\tau, (1-\tau)\varepsilon)$-center.*

Proof. 1. Given a $(1, \varepsilon)$-centered set $X \subseteq \{-1,1\}^n$, write a linear program in variables $z_1, \ldots, z_n \in \mathbb{R}$ with the following equations

$$\forall x \in X, \quad \sum_i x_i z_i \geq \varepsilon n; \qquad \forall i \in [n], \quad -1 \leq z_i \leq 1$$

The solution will be some $z \in [-1,1]^n$, and output $\tilde{z}$ the randomized rounding of z, i.e. $\tilde{z}_i = 1$ with probability $(1+z_i)/2$.

A standard Chernoff bound will show that $|\langle \tilde{z}, x\rangle - \langle z, x\rangle| < \sqrt{2\log|X|\,/n}$ for all $x \in X$ with high probability.

2. Given a $(1-\nu, \varepsilon)$-centered set X we write a similar linear program, except we add 'violation' variables v_x per each x as follows

$$\forall i \in [n] \quad -1 \leq z_i \leq 1$$

$$\forall x \in X \quad 0 \leq v_x \leq 1 + \varepsilon$$

$$\forall x \in X \quad \frac{1}{n}\sum_i x_i z_i + v_x \geq \varepsilon$$

and then we find a solution minimizing $val = \frac{1}{|X|}\sum_x v_x$. Again, the final output of the algorithm is a randomized rounding $\tilde{z}$ of the solution z.

It is easy to see that the solution $z = \bar{0}$ with $\forall x,\ v_x = \varepsilon$ is a feasible solution whose value is $val = \varepsilon$. A more interesting solution is where z is the promised $(1-\nu, \varepsilon)$-center, and for

every equation violated by x outside this ball, we set $v_x = 1 + \varepsilon$. This solution has value $val = 0 \cdot (1 - \nu) + (1 + \varepsilon) \cdot \nu = (1 + \varepsilon)\nu$. These two solutions show that the solution to the linear program gives information only as long as $(1 + \varepsilon)\nu < \varepsilon$. Suppose $z, \{v_x\}_{x \in X}$ is the solution for this system, with value $v = \mathbb{E}_x[v_x] \leq (1 + \varepsilon)\nu < \varepsilon$. By Markov's inequality, at most $1/a$ fraction of the x's have $v_x > av$. The remaining $1 - 1/a$ equations are satisfied to within av, as claimed. The last conclusion follows by setting $a = 1/\tau$.

■

3.2 Clustering with Few Outliers: the list decoding regime

We now turn to the clustering question when the input consists of mostly noise. In other words, where the data set X is guaranteed to be (δ, ε)-centered, for values of δ, ε as small as $1/\text{poly}\log(n)$.

Clearly, X might have several distinct (δ, ε)-centers, and ideally we would like an algorithm that outputs a list of all of them. To control the length of the list, we must settle for a list of centers that 'represent' all the possible centers in X. One natural way to define 'represent' is by saying that z represents z' if the symmetric difference between $C_\varepsilon(z)$ and $C_\varepsilon(z')$ is small (compared to their size). However, this notion turns out to be insufficient. It is easy to describe a set X of points that are highly correlated to a single center, and yet could be explained by an exponential number of other centers, whose pairwise symmetric difference is large, see Section 5 for details. This example shows that the best we can hope for is an algorithm that outputs a list of clusters that "approximately cover" every cluster, in the sense that every (δ, ε) cluster is guaranteed to be approximately contained in a cluster from the list.

Definition 3.1. Let $\tau > 0$. We say that $A \subseteq_\tau B$ if $|A \setminus B| \leq \tau |A|$.

Definition 3.2 (τ-maximal center)**.** For a set X, and parameters $\delta, \varepsilon, \tau > 0$ we say that a (δ, ε)-center z is τ-maximal if the following holds: For every $y \in \{-1, 1\}^n$, if $C_\varepsilon(z) \subseteq_\tau C_{(1-\tau)\varepsilon}(y)$, then

$$\left|C_{(1-\tau)\varepsilon}(y)\right| < (1 + \tau) \left|C_\varepsilon(z)\right|.$$

This definition says that if there is some y whose cluster approximately contains the cluster of a maximal z, then the cluster of y is not much larger. With this definition, we can now state the formal version of Theorem 1.4:

Theorem 3.3 (Formal Version of Theorem 1.4)**.** *Let $\varepsilon, \delta, \tau > 0$ be parameters. Let $X \subset \{-1, 1\}^n$ be an N-element set that is (δ, ε)-centered. There exists an algorithm that runs in time polynomial in $n, N, 2^{O(\frac{1}{\varepsilon^2 \tau^2} \log 1/\tau^2 \delta\varepsilon)}$, and outputs a list L of points each in $\{-1, 1\}^n$, such that with probability $1 - 2^{-n}$ the following holds:*

1. *Each $y \in L$ is a $((1 - \tau) \cdot \delta, (1 - 2\tau) \cdot \varepsilon)$ center for X.*
2. *For each $z \in \{-1, 1\}^n$ which is a (δ, ε)-center there exists $y \in L$ such that,*

$$C_\varepsilon(z) \subseteq_{3\tau/\delta} C_{(1-2\tau)\varepsilon}(y).$$

3. *Moreover, if z is $3\tau/\delta$-maximal, then*

$$\left|C_\varepsilon(z) \triangle C_{(1-2\tau)\varepsilon}(y)\right| < \frac{6\tau}{\delta} \left|C_\varepsilon(z)\right|.$$

The following lemma is the main technical tool in the proof:

Lemma 3.4. *Let $\varepsilon, \delta, \tau > 0$ be parameters. Let $X \subset \{-1,1\}^n$ be an N-element set that is (δ, ε)-centered. There exists an algorithm that runs in time $\mathrm{poly}(n, N, 2^k)$, where $k = O(\frac{1}{\varepsilon^2\tau^2} \log 1/\tau^2\delta\varepsilon)$, and outputs a list L of at most 2^k points each in $\{-1,1\}^n$, such that:*

- *Each $y \in L$ is a $((1-\tau)\cdot\delta, (1-2\tau)\cdot\varepsilon)$ center for X.*
- *For each $z \in \{-1,1\}^n$ which is a (δ,ε)-center with probability $> 1/2$ there exists $y \in L$ such that,*

$$C_\varepsilon(z) \subseteq_{3\tau/\delta} C_{(1-2\tau)\varepsilon}(y).$$

Proof of Theorem 3.3 using Lemma 3.4. We apply the algorithm implied by Lemma 3.4 $t = 2n$ iterations, and concatenate the list obtained in each iteration. Clearly, the first item holds, namely each element in the list is a $(1-\tau)\delta, (1-2\tau)\varepsilon)$-center. We proceed to prove Item 2. We say that a center z is discovered in the i-th iteration, if there exists y in the list produced in the i-th iteration such that:

$$C_\varepsilon(z) \subseteq_{3\tau/\delta} C_{(1-2\tau)\varepsilon}(y).$$

Fix a (δ,ε)-center z, by Lemma 3.4 we get that the probability that z is discovered in the i-th iteration is at least $1/2$, so the probability that it remains undiscovered after t iterations is at most $1/2^t < 2^{-2n}$. Taking union bound on all centers the probability that there exits a center that remains undiscovered is at most $1/2^n$, and this completes the proof of Item 2 of the theorem.

For Item 3, let z be a $3\tau/\delta$-maximal center. By Item 2 we get that there exists $y \in L$ such that:

$$C_\varepsilon(z) \subset_{3\tau/\delta} C_{(1-2\tau)\varepsilon}(y).$$

In order to complete the proof we show that opposite difference is also small. Since z is $3\tau/\delta$-maximal, we get that: $\left|C_{(1-2\tau)\varepsilon}(y)\right| < (1+3\tau/\delta)\left|C_\varepsilon(z)\right|$, so:

$$\begin{aligned}\left|C_{(1-2\tau)\varepsilon}(y) \setminus C_\varepsilon(z)\right| &= \left|C_{(1-2\tau)\varepsilon}(y)\right| - \left|C_{(1-2\tau)\varepsilon}(y) \cap C_\varepsilon(z)\right| \\ &\leq (1+3\tau/\delta)\left|C_\varepsilon(z)\right| - (1-3\tau/\delta)\left|C_\varepsilon(z)\right| \\ &\leq 6\tau/\delta\left|C_\varepsilon(z)\right|.\end{aligned}$$

■

Now we turn to prove Lemma 3.4:

Proof of Lemma 3.4. The proof of this lemma follows by randomly restricting the points to a smaller dimensional space and then enumerating to find a good approximation for the cluster. The approximate center is then found by applying Theorem 1.3 on the approximate cluster. This algorithm is relatively efficient when $\varepsilon, \delta = \mathrm{poly}(1/\log n)$. The algorithm is as follows.

Algorithm 1 Randomly Restrict and Enumerate

Input: A (δ,ε)-centered set X.
Parameters: $k = \frac{C}{\varepsilon^2\tau^2} \log 1/\tau^2\delta\varepsilon$ for some large enough C to be determined later, and $\tau > 0$.

1. Choose at random a multiset $K \subseteq [n]$ by selecting a random $i \in [n]$ into K repeatedly k times with replacement.
2. For each $y \in \{-1,1\}^k$ let $X(y) = \{x \in X \mid \langle x_K, y\rangle \geq (1-\tau/2)\cdot\varepsilon\}$, and compute the center $z(y)$ of $X(y)$ using the linear programming algorithm from Item 3 of Theorem 1.3 with correlation parameter $(1-\tau)\varepsilon$. If $z(y)$ is a $((1-\tau)\cdot\delta, (1-2\tau)\cdot\varepsilon)$ center for X then output it.

Clearly, each center produced by the list is a $((1-\tau)\cdot\delta,(1-2\tau)\cdot\varepsilon)$ center for X. Moreover, the list size is bounded by 2^k. It is left to prove the second and the third item of the lemma. Let z^* be a (δ,ε)-center, and consider the set

$$X^* := \{x \in X \mid \langle x_K, z_K^*\rangle > \varepsilon(1-\tau/2)\}\,.$$

We will prove that $1-\gamma$ fraction of the elements of X^* also belong to $C_{(1-\tau)\varepsilon}(z^*)$, which means that X^* is $(1-\gamma,(1-\tau)\varepsilon)$-centered. This implies that at step 2 our algorithm will output some center z' of X^* (because $X^* = X(y)$ for $y = z_K^*$). We will then prove that $C_\varepsilon(z^*) \subset_{3\tau/\delta} C_{(1-2\tau)\varepsilon}(z')$ which means that z^* is covered by our list.

We first claim that the sampling is good enough.

Claim 3.5. *We say that $x \in X$ is typical with respect to K if $|\langle x, z^*\rangle - \langle x_K, z_K^*\rangle| \le \frac{\tau\varepsilon}{2}$. Then for at least half of the choices of K, the fraction of typical x's is at least $1-\gamma$ for $\gamma := 2\exp(-\tau^2\varepsilon^2 k/12) < \tau^2\delta\varepsilon/8$.*

Proof. We first show that $\Pr_{x\in X,K}[x \text{ is not typical}] \le \gamma$. In fact, we show this for each fixed x separately. For a random i, $x_i z_i^*$ can be viewed as a random ± 1 variable whose expectation is $\langle x, z^*\rangle$. By a Chernoff bound the probability that $|\langle x, z^*\rangle - \langle x_K, z_K^*\rangle| > \tau\varepsilon/2$ is at most $\gamma/2$.

By an averaging argument this means that for at least half of the choices of K have no more than γ atypical x's, which gives the claim. ■

Suppose from now on that K is as in the claim, so there are at most γ atypical points $x \in X$. Clearly, every point in X^* for which $\langle x, z^*\rangle < (1-\tau)\varepsilon$ must be atypical, and every typical point in $C_\varepsilon(z)$ is also in X^*. So z^* is a $(1-\gamma/(\delta-\gamma),(1-\tau)\varepsilon)$-center for this set. These conditions allow step 2 of the algorithm to work, according to Theorem 1.3, and to output a point z' that is a $(1-\tau,(1-2\tau)\varepsilon)$-center for X^*. The conditions of Theorem 1.3 are simply that $\gamma/(\delta-\gamma) < 2\gamma/\delta < \tau^2\frac{\varepsilon}{1-\varepsilon}$ which clearly holds by the choice of k.

In order to complete the proof, we show

$$\left|C_\varepsilon(z^*) \setminus C_{(1-2\tau)\varepsilon}(z')\right| \le \gamma\,|X| + \frac{\tau}{\delta-\gamma}\,|X^*| \le \frac{3\tau}{\delta}\,|C_\varepsilon(z^*)|$$

This is simply since the measure of points that are in $C_\varepsilon(z^*)$ but not in X^* is at most $\gamma\,|X|$, and the measure of points in X^* but not in $C_{(1-2\tau)\varepsilon}(z')$ is at most $\tau\,|X^*|$ (by the guarantee of Theorem 1.3), so the measure of points in $C_\varepsilon(z^*)\cap X^*$ which are not in $C_{(1-2\tau)\varepsilon}(z')$ is at most $\tau/(\delta-\gamma)$. For the last inequality we rely on the fact that $|C_\varepsilon(z^*)| \ge \delta$ and $\gamma/\delta < \tau$.

■

3.3 Approximating very small margins (proof of Theorem 1.5)

In our next theorem, we consider much smaller margins, say $\varepsilon = 1/n^{0.1}$. Here enumerating over a space of dimension $1/\varepsilon$ is out of the question. Instead, our argument uses the Johnson bound to deduce that one of the points of X is already a good "approximate" center.

Theorem 1.5. *Let $X \subset \{-1,1\}^n$ be an (δ,ε)-centered set of N elements and let $\varepsilon \gg \log n/\sqrt{n}$. There is an algorithm that runs in time* $\mathrm{poly}(N)$ *and outputs a list L such that:*

- *Each $z \in L$ is a $(2\delta\varepsilon,\varepsilon^2)$-center.*

- *For every center z, there exists a center $z' \in L$, such that at least 2ε fraction of the points $x \in X$ satisfying $\langle x, z\rangle > \varepsilon$ are also satisfying $\langle x, z'\rangle > \varepsilon^2$ (We call z' an approximate center for z).*

In order to prove the theorem we prove first the following lemma:

Lemma 3.6. *Let $\varepsilon > 0$ and let $X \subseteq \{-1,1\}^n$ be any $(1,\varepsilon)$-centered set. Then, there exists $x \in X$ such that x is a $(2\varepsilon, \varepsilon^2)$ for X.*

Proof. Let $X = \{x_1, \ldots, x_N\}$ be an $(1,\varepsilon)$-centered set, let z be an $(1,\varepsilon)$-center. Let $X' \subseteq X$ constructed by adding x_i into X' if for all $j < i$ we have $\langle x_i, x_j\rangle < \varepsilon^2$.

X' can be viewed as an error correcting code with distance at least $1/2 - \varepsilon^2$ and hence by Lemma 2.1 a ball of radius $1/2 - \varepsilon$ around z cannot contain more than $1/2\varepsilon$ of them, so $|X'| \leq 1/2\varepsilon$.

Now, for every $x' \in X'$ let $p(x') = \Pr_{x\in X}[\langle x, x'\rangle \geq \varepsilon^2]$. Observe that for each $x \in X$ there exists $x' \in X'$ such that $\langle x, x'\rangle > \varepsilon^2$, so $\sum_{x'\in X'} p(x') \geq 1$. Therefore, by averaging argument, there exists $x' \in X'$ with $p(x') \geq 1/|X'| \geq 2\varepsilon$. Clearly, this point x' is a $(2\varepsilon, \varepsilon^2)$-center for X and the proof is done. ■

Proof of Theorem 1.5: The algorithm is as follows: For each $y \in X$ we include $y \in L$ if it is a $(2\delta\varepsilon, \varepsilon^2)$-center for X. Clearly, the first item of the theorem holds. To prove the second part, we observe that by Claim 3.6, for each (δ,ε)-center z, one of the points in X is an approximate center for z, as required. ■

4 Hardness of Approximating the gap-Clustering Problem

In this section we study the hardness of the task of finding a (δ,ε) over various choices of parameters. We first show that it is infeasible to solve the task of finding a (δ,ε)-center without approximation, even when $\delta = 1$.

Claim 4.1. *Unless $\mathcal{BPP} \supseteq$ **NP**, given a $(1,\varepsilon)$-centered set it is infeasible to find a $(1,\varepsilon)$-center, for $\varepsilon = n^{-1/3}(1 - o(1))$.*

Proof Sketch. We do not give a whole proof, but rather sketch the proof. The proof of the completeness and soundness resembles the proof of Theorem 1.7. We reduce the 3SAT problem into our problem as follows: Given a 3SAT formula ψ with n variables and m constraints, we translate it into $(1,\varepsilon)$ clustering problem instance $X \subseteq \{-1,1\}^{(n+2)r}$ with $|X| = m$, where $r = n^2$.

Each variable y_i in ψ is represented by r coordinates and we index the coordinates of a point $x \in X \subseteq \{-1,1\}^{nr}$ by a double index $x_{i,s}$ for $i \in [n]$ and $s \in [r]$. Each clause $\ell_i \vee \ell_j \vee \ell_k$ gives rise to the following point:

$$x_{m,s} = \begin{cases} (-1)^{sign(\ell_m)+1} & \text{if } m \in \{i,j,k\}, \\ 1 & \text{if } m \in \{n+1, n+2\}, \\ \text{coinflip} & \text{otherwise} \end{cases}$$

It is not hard to see that if ψ is satisfied if and only if X has a $(1, \frac{1}{n}(1 - o(1)))$-center. ■

We remark that by adding random points to X the δ parameter can be made smaller than 1. Additionally, by adding dummy coordinates to all the points in X, all containing the value 1, the ε parameter can be made to approach 1.

Of course, the more interesting question is that of approximate hardness. We show that one cannot even find an approximate center when such exists. Recall the definition the problem (Definition 1.6), where the task is given a set of points X distinguish between the case that there exists a (δ, ε)-center, and the case where no (δ', ε')-center exists.

There are two key points which we like to address in the parameters settings: First we would like to get as large as possible a gap between δ and δ' and ε and ε'. The second is locating the gap-finding the largest ε and δ for which the problem is hard. By the following claim, the larger ε and δ are, the stronger is the hardness result.

Claim 4.2. *Assume $\varepsilon, \varepsilon' \gg \log N/\sqrt{n}$, then:*

- *For every $c > 1$, a hardness result with parameters $(\delta, \varepsilon, \delta', \varepsilon')$, implies a hardness result with parameters $(\delta/c, \varepsilon, \delta'/c, \varepsilon')$.*
- *For every $c > 1$, a hardness result with parameters $(\delta, \varepsilon, \delta', \varepsilon')$, implies a hardness result with parameters $(\delta, \varepsilon/c, \delta', \varepsilon'/c)$.*

Proof. The first item holds since we can add random points into the original set X. Note that with overwhelming probability each random point s satisfies $\langle s, x \rangle < \varepsilon'$ for every $x \in X$. The second item holds since we can add new coordinates to the original instance, by setting a random value in each of the new coordinates. ■

There are four parameters involved so the gaps and tradeoffs between the parameters can become quite complicated. We separate the task of finding the largest δ-gap and the task of finding the largest ε-gap.

Our first hardness result shows a factor 2 gap between ε and ε'. It shows that it is hard to distinguish between the case that there exists a (δ, ε)-center, and the case that there is no $(\delta/c, \varepsilon/2)$ for some constant $c > 1$, ε which is an arbitrarily large polynomial in $1/n$, and δ that is a constant. This is proved in Section 4.1.

Our next result shows a polynomial gap between δ and δ'. It shows that it is hard to distinguish between the case that there exists a (δ, ε)-center, and the case that no $(\frac{\delta}{n^c}, \varepsilon')$-center exists, for c being some constant, ε, δ which are poly$(1/n)$, and $\varepsilon' = (1 - o(1))\varepsilon$. This is proved in Section 4.2.

The starting point for our reductions for both Theorem 1.7 and Theorem 1.8 is the MAX-DICUT problem: Given n Boolean variables $y_1, \ldots, y_n$, and m constraints of the form $\neg y_i \wedge y_j$, the goal is to satisfy the maximal number of constraints. Now we state the gap version of the **NP**-hardness result for MAX-DICUT obtained by [TSSW96].

Theorem 4.3 ([TSSW96])**.** *There exists a constant $\gamma > 0$ such that given a MAX-DICUT instance $\mathcal{I}$ with n variables, it is* **NP***-hard to decide whether there is an assignment that satisfies γ fraction of the constraints or that every assignment satisfies at most $\frac{12}{13}\gamma$ fraction of the constraints.*

Moreover, every variable y_i appears in at most d/n-fraction of the constraints, where d is some constant.

The following definition would be useful in the proof:

Definition 4.4. The generalized gap-clustering problem with parameters $(\delta, \varepsilon, \delta', \varepsilon')$:

The input of the problem is a set of point $X \subset \{-1, 0, 1\}^n$. The goal is to distinguish between the following cases:

- There exists a (δ, ε)-center for X.
- There is no (δ', ε')-center for X.

The following lemma asserts that if the generalized gap-clustering problem is hard, then the gap-clustering is hard with essentially the same sets of parameters.

Lemma 4.5. *There exists a randomized algorithm that takes an instance for the generalized gap-clustering problem X, and outputs an instance for the gap-clustering problem X' such that with probability at least $9/10$ the following holds: If X has (δ,ε)-center, then X' has $(\delta-\eta,\varepsilon-\zeta)$-center. If X has no (δ',ε')-center, then X' has no $(\delta'+\eta,\varepsilon'+\zeta)$-center, for any value of $\zeta > 0$ and $\eta > 2\exp^{-\Omega(\zeta^2 n)}$. The running time of the algorithm is* $\mathrm{poly}(1/\eta, n, N)$.

Proof. Given an instance $X \subset \{-1,0,1\}$, $|X| = N$ for generalized gap-clustering we translate it into a set $X' \subset \{-1,1\}^n$, $|X'| = tN$, where $t = \Omega(n\log(N)/\eta^2)$. Each point $x \in X$ gives rise to t points sampled independently at random from the following distribution:

$$x'_i = \begin{cases} x_i & \text{if } x_i \in \{-1,1\}, \\ \text{coin flip} & \text{otherwise} \end{cases}$$

where each coinflip coordinate in x is drawn independently and uniformly at random in $\{-1,1\}$. We now prove completeness and soundness:

Completeness: If there exists a (δ,ε)-center z for X, then we show that with probability at least $9/10$, z is a $(\delta-\eta,\varepsilon-\zeta)$-center for X'.

Let $x \in X$ satisfying $\langle z,x\rangle > \varepsilon$. Let S_x be the set of coordinates in x for which $x_i = 0$. For each $i \in S_x$, $x'_i \cdot z_i$ is a ± 1 random variable with expectation 0, so $\mathbb{E}\langle z,x'\rangle = \langle z,x\rangle$. It is easy to see that the larger the cardinality of S_x is, the larger is the probability that $\langle z,x'\rangle < \langle z,x\rangle - \zeta$. Now, taking S_x to be maximal (i.e. $[n]$), by Chernoff bound the probability that $\langle z,x'\rangle < \varepsilon-\zeta$ is at most $\exp^{-\Omega(\zeta^2 n)} < \eta/2$. By Chernoff bound again, the probability that there would be more than η points associated with x, that satisfy $\langle z,x'\rangle < \varepsilon-\zeta$ is bounded by $\exp^{-\Omega(\eta^2 t)} < 1/10m$. Taking union bound on all points in X satisfying $\langle z,x\rangle > \varepsilon$, the completeness follows.

Soundness: If X has no (δ',ε')-center, then we show that with probability at least $9/10$, there is no $(\delta'+\eta,\varepsilon'+\zeta)$ for X'.

Take $z \in \{-1,1\}^n$, and take x such that $\langle z,x\rangle < \varepsilon'$. Analogously to the completeness case, the probability that x' associated with x satisfies $\langle x',z\rangle > \varepsilon'+\zeta$ is bounded by $\exp^{-\Omega(\zeta^2 n)} < \eta/2$. Therefore, fix z and x, the probability that more than $\delta'+\eta$ points $x' \in X'$ that associated with x satisfy $\langle z,x'\rangle > \varepsilon'+\zeta$ is bounded by $\exp^{-\Omega(\eta^2 t)} < \frac{1}{10N2^n}$. By taking union bound on all centers and points in X, the soundness follows. ■

4.1 Maximizing ε-gap for the gap-Clustering Problem

We begin with the following theorem, a hardness of approximating the gap-clustering problem that addresses the ϵ ratio, showing that it is hard to decide if a given set has a $(\delta, 2\varepsilon)$ center and (roughly) a (δ,ε) one.

Theorem 1.7. *Unless $\mathcal{BPP} \supseteq$ **NP**, there exist constants $\delta, c > 1$, such that for every constant $\alpha > 2$, it is infeasible to solve the gap-clustering problem with parameters: $\delta, \delta' = \delta/c$, $\varepsilon = \frac{2}{n^{1/\alpha}} - o(\frac{1}{n^{1/\alpha}})$ and $\varepsilon' = \frac{1}{n^{1/\alpha}} + \omega(\frac{1}{n^{1/\alpha}})$.*

The following lemma shows that there is a polynomial time reduction that given an instance for the gap version of MAX-DICUT produces an instance for the generalized gap-clustering problem.

Lemma 4.6. *There exists a constant γ, and for every constant value of $\eta > 0$ and $\alpha > 2$, there exists a polynomial time algorithm that when given MAX-DICUT instance $\mathcal{I}$ with n variables and*

m constraints as an input, produces an instance X for the generalized gap-clustering problem, with $|X| = m$ points, each lies in $\{-1,1\}^{\tilde{n}}$, where $\tilde{n} = n^{\alpha}$ such that:

- *(Completeness:) If $\mathcal{I}$ has an assignment that satisfies at least γ fraction of the constraints, then, there exists a (γ, ε)-center for X, for $\varepsilon = \frac{2}{n} = \frac{2}{\tilde{n}^{1/\alpha}}$.*
- *(Soundness:) If every assignment for $\mathcal{I}$ satisfies at most $\frac{12}{13}\gamma$ fraction of the constraints, then, with probability larger than $2/3$, there is no $(\frac{12}{13}\gamma, \varepsilon/2)$-center for X.*

Proof of Theorem 1.7 using Lemma 4.6 and Lemma 4.5. We take an instance $\mathcal{I}$ for MAX-DICUT with n variables, such that either at least γ-fraction of the constraints are satisfied or at most $\frac{12}{13}\gamma$-fraction of the constraints are satisfied. We first translate it using Lemma 4.6 into an instance $X \in \{-1,0,1\}^{\tilde{n}}$, $\tilde{n} = n^{\alpha}$, for the generalized gap-clustering problem such that YES-instances have $(\gamma, \frac{2}{\tilde{n}^{1/\alpha}})$-center, while NO-instances have no $(\frac{12}{13}\gamma, \frac{1}{\tilde{n}^{1/\alpha}})$-center.

Then we apply Lemma 4.5, with $\eta < \frac{12}{39}\gamma$, and $\zeta = O(1/\sqrt{\tilde{n}}) = o(1/n^{1/\alpha})$ (observe that $\eta > 2\exp^{-(\zeta^2\tilde{n})}$), and get that YES-instances are translated into X' that have $(\gamma - \eta, \varepsilon - \zeta)$-center, while NO-instances have no $(\frac{12}{13}\gamma + \eta, \varepsilon' + \zeta)$-center, the Theorem follows.

■

We now prove Lemma 4.6:

Proof of Lemma 4.6. Given a MAX-DICUT instance $\mathcal{I}$ with n variables and m constraints, we translate it into $(\delta, \varepsilon, \delta', \varepsilon')$ gap-clustering problem instance $X \subseteq \{-1,1\}^{nr}$ with $|X| = m$, where $r = n^{\alpha-1}$.

Each variable y_i is represented by r coordinates and we index the coordinates of a point $x \in X \subseteq \{-1,1\}^{nr}$ by a double index $x_{i,s}$ for $i \in [n]$ and $s \in [r]$. Each constraint $c_{i,j} = \neg y_i \wedge y_j$ gives rise to the following point:

$$x_{k,s} = \begin{cases} -1 & \text{if } k = i, \\ 1 & \text{if } k = j, \\ 0 & \text{otherwise} \end{cases}$$

We now prove completeness and soundness:

Completeness: Assuming there exists an assignment a that satisfies at least γ fraction of the constraints of $\mathcal{I}$, we show that there exists a (γ, ε)-center in X. Consider the point $z \in \{-1,1\}^{\tilde{n}}$ defined as follows: for all $s \in [r], i \in [n]$, set $z_{i,s} = (-1)^{(a(y_i)+1)}$. Then, for every constraint $c_{i,j}$ satisfied by a, it holds that $\sum_{s\in[r]} -z_{s,i} + z_{j,s} = 2r$. Therefore, if x is associated with $c_{i,j}$ and $c_{i,j}$ is satisfied by a, we get: $\langle x, z\rangle = 2r/rn = 2/n$, and the completeness follows.

Soundness: Now take an instance $\mathcal{I}$ for which there is no assignment that satisfies more than $\frac{12}{13}\gamma$ fraction of constraints. We show that there is no $(\frac{12}{13}\gamma, 1/n)$-center in the random instance X.

Let $z \in \{-1,1\}^{\tilde{n}}$, we define an assignment a for $y_1, \ldots, y_n$ assigning for each $i \in [n]$ the value $a(y_i) = (-1)^{(\mathrm{MAJ}\{z_{i,s}\}_{s\in[r]}+1)}$. We show that if z is a $(\frac{12}{13}\gamma, 1/n)$-center for X, then a satisfies more than $\frac{12}{13}\gamma$ fraction of the constraints of $\mathcal{I}$, contradicting the soundness assumption.

For any constraint $c_{i,j}$, consider the point x associated with $c_{i,j}$. Observe that $\langle z, x\rangle > 1/n$ iff $\sum_{s\in[r]}(-z_{i,s} + z_{j,s}) > r$. In such case the majority value of the $z_{i,s}$ is -1, and the majority of the $z_{j,s}$ is 1. Therefore, $c_{i,j}$ is satisfied by a, and the proof is done.

■

4.2 Maximizing δ-gap for the gap-Clustering Problem

Next, we show a polynomial gap between δ and δ'.

Theorem 1.8. *Unless $\mathcal{BPP} \supseteq$ **NP**, there exist constants $c_1 > 0, c_2 > 0$ such that it is infeasible to solve the gap-clustering problem with parameters: $\delta = n^{-c_1}, \delta' = \frac{\delta}{n^{c_2}}$ and $\varepsilon > \varepsilon' = \Theta(n^{-1/3})$.*

The main step of the proof is the following lemma which we prove next.

Lemma 4.7. *There exists a constant c, such that for $\ell = c \log n$, and $\beta = 4\left(\frac{12}{13}\right)^{\ell}$ there exists a probabilistic polynomial time algorithm that when given a regular MAX-DICUT instance $\mathcal{I}$ with n variables and m constraints as an input, produces an instance X for generalized gap-clustering problem, with $N = \Theta\left(n\left(\frac{13}{12\gamma}\right)^{2\ell}\right)$ points, each lies in $\{-1,1\}^{\tilde{n}}$, $\tilde{n} = n^3$, such that:*

- *(Completeness:) If $\mathcal{I}$ has an assignment that satisfies at least γ fraction of the constraints, then, with probability larger than $2/3$, there exists a (δ, ε)-center for X, for $\delta = \gamma^{\ell}/2$ and $\varepsilon = \frac{2\ell}{n}$.*
- *(Soundness:) If every assignment for $\mathcal{I}$ satisfies at most $\frac{12}{13}\gamma$ fraction of the constraints, then, with probability larger than $2/3$, there is no $(\beta \cdot \delta, \varepsilon')$-center for X, for $\varepsilon' = \frac{2\ell-1}{n}$.*

Proof of Lemma 4.7. Our first step is to take the instance $\mathcal{I}$ and translate it into generalized gap-clustering problem X, such that each point lies in $\{-1,0,1\}^{\tilde{n}}$, where $\tilde{n} = n^3$, as done in Lemma 4.6. We get that if $\mathcal{I}$ is a YES-instance then X has $(\gamma, 2/n)$-center, while if $\mathcal{I}$ is a NO-instance then X that has no $(\frac{12}{13}\gamma, 1/n)$-center. Our next step is to associate for each $\ell = c\log(n)$-tuple of points $x_1, \ldots, x_\ell$, a new point $x = x_1 + \cdots + x_\ell$.

There are m^ℓ ℓ-tuples, so we cannot take all the ℓ tuples. Therefore, we take only a random subset of them. We must be careful to take point only points with no common non-zero entries so that the sum of the points is still in $\{-1,0,1\}^n$. Note that since in $\mathcal{I}$ every variable appears in d/n fraction of the constraints, then the fraction of ℓ tuples that have common no-zero entry is bounded by $d\ell^2/n$.

Formally, we choose at random a subset S of non intersecting ℓ points $x_1, \ldots, x_\ell$, for $|S| > \Omega\left(n\left(\frac{13}{12\gamma}\right)^{2\ell}\right)$. Each such ℓ tuple gives rise to a point $x = x_1 + \cdots + x_\ell$ in X'. Now we prove completeness and soundness.

Completeness: Consider an instance $\mathcal{I}$ for MAX-DICUT for which there exists an assignment a that satisfies at least γ fraction of the constraints. In such a case there exits a $(\gamma, 2/n)$-center z for X. Since we take only non-intersecting points then the probability that all them are ε-correlated to z is at least $\gamma^\ell - d\ell^2/n > \gamma^\ell/2$. Using Chernoff inequality, the probability that we sample less than $\gamma^\ell/4$ such tuples is bounded by $\exp^{-\Omega(\gamma^{2\ell}|S|)}$ which is at most $1/3$ by our choice of parameters.

Let $x_1, \ldots, x_\ell$ be an ℓ tuple of non-intersecting points, which are all $2/n$-correlated to z, then $x = x_1 + \ldots x_\ell$ satisfies: $\langle x, z\rangle > \frac{2\ell}{n}$. We get that with probability at least $2/3$ at least $\gamma^\ell/4$-fraction of the points in X' are $\frac{2\ell}{n}$-correlated to z, and the completeness follows.

Soundness: Consider an instance $\mathcal{I}$ for MAX-DICUT for which every assignment a satisfies at most $\frac{12}{13}\gamma$ fraction of the constraints. In such a case there there is no $(\frac{12}{13}\gamma, 1/n)$-center z for X. Fix $z \in \{-1,1\}^n$, and let $x \in X'$, $x = x_1 + \cdots + x_\ell$. Note that whenever $\langle x, z\rangle > \frac{2\ell-1}{n}$, then it holds that for all $i \in [\ell]$, $\langle x_i, z\rangle > 1/n$ (this is true since for every point $x \in X$ the fraction of non-zero coordinates is $2/n$ so $\langle x, z\rangle \leq 2/n$).

If we choose ℓ points at random, the the probability that each of them is $1/n$-correlated to z is bounded by $(\frac{12}{13}\gamma)^\ell$. The probability that there are more than $2(\frac{12}{13}\gamma)^\ell$-fraction of such tuples in S is bounded by $\exp^{-\Omega((\frac{12}{13}\gamma)^{2\ell}|S|)}$ which is at most $\frac{1}{10\cdot 2^{\tilde{n}}}$ by our choice of parameters. Therefore, taking a union bound on the set of all possible centers the probability that there exists a center for which more than $2(\frac{12}{13}\gamma)^\ell$ of the tuples are satisfying that all points are $1/n$-correlated to z is bounded by $1/3$. If that is not the case, then there is no $(2(\frac{12}{13}\gamma)^\ell = 4\delta/\beta, \frac{2\ell-1}{n})$-center for X', the soundness follows.

■

We conclude with the proof of Theorem 1.8:

Proof of Theorem 1.8 Using Lemma 4.7. We take an instance $\mathcal{I}$ for MAX-DICUT with n variables, such that either at least γ-fraction of the constraints are satisfied or at most $\frac{12}{13}\gamma$-fraction of the constraints are satisfied. We first translate it using Lemma 4.7 into an instance $X \subset \{-1,0,1\}^{\tilde{n}}$ for the generalized gap-clustering problem such that YES-instances have $(\gamma^\ell/2, \frac{2\ell}{n})$-center, while NO-instances have no $(2\left(\frac{12}{13}\gamma\right)^\ell, \frac{2\ell-1}{n})$-center.

Then we apply Lemma 4.5, with $\eta < (\frac{12}{13}\gamma)^\ell/16$, and $\zeta = 1/3n$ (observe that $\eta > 2\exp^{-\Omega(\zeta^2\tilde{n})}$). We get that YES-instances are translated into X' that have $(\gamma^\ell/2-\eta, \varepsilon-\zeta)$-center, while NO-instances have no $(2\left(\frac{12}{13}\gamma\right)^\ell+\eta, \varepsilon'+\zeta)$-center. Consider the ratio $\frac{\gamma^\ell/2-\eta}{2\left(\frac{12}{13}\gamma\right)^\ell+\eta} > \frac{\gamma^\ell/4}{4\left(\frac{12}{13}\gamma\right)^\ell} = \left(\frac{13}{12\gamma}\right)^\ell/16$, which is polynomial in n, the Theorem follows.

■

A similar proof can be used to show:

Lemma 4.8. *For every constant $c > 0$, there exist constants $c_1, c_2 > 0$ such that unless $\mathcal{BPP} \supseteq \mathbf{NP}$, it is infeasible to solve the gap-clustering problem with parameters: $\delta = c_1, \delta' = c_1/c$ and $\varepsilon' = c_2\varepsilon$, and ε which is $\Theta(n^{-1/3})$.*

The proof of Lemma 4.8 follows by the same proof of Theorem 1.8, and we omit it.

5 Bounding the Centers List Size

In this section we study the information-theoretic bound to the list size of all possible centers, providing the best-possible algorithms that attempts to list all the (δ,ε)-centers.

Our first result shows that when we consider a list of (δ,ε)-centers, such that for every pair of centers in the list, their clusters have a large symmetric difference, then it might be the case that the list size is exponential in n. These centers do not have the property of being τ-maximal (see definition 3.2), so this shows that the requirement of being τ-maximal in Theorem 3.3 is essential.

Lemma 5.1. *For any $\varepsilon > 0$ and $0 < \delta < 1/2$, there exists a set X, and a t-size list L of points, $t = 2^{\Omega(n)}$, such that:*

- *For each $z \in L$, z is a $(1/2-\varepsilon,\varepsilon)$-center for X.*
- *For $z \neq z' \in L$: $|C_\varepsilon(z) \triangle C_\varepsilon(z')| > 2\delta X$ (where $A \triangle B$ is the symmetric difference between A and B).*

Proof. We construct a set X of vectors that has a global $(1, 1-1/n)$-center, yet at the same time has an exponential number of $(1/2-\varepsilon, \varepsilon)$ centers $z^{(1)}, z^{(2)}, \ldots$ such that the clusters of $z^{(i)}$ and $z^{(j)}$ have large symmetric difference if $i \neq j$. This means that there is no list of clusters of polynomial size that is close to *all* non-trivial centers of X.

Denote e_i the vector with 1 in the i-th coordinate, and -1 in the rest. Let $X = \{e_i \mid i \in [n]\}$. Clearly the $-\mathbf{1}$ vector is a $(1, 1-1/n)$-center for this set.

Let $S \subset [n]$ have size t and let z be -1 on coordinates in S and 1 otherwise. We have:

$$\langle z, e_i \rangle = \begin{cases} |S| - \left|\bar{S}\right| + 2 = t - (n-t) + 2 = 2t - n + 2 & \text{if } i \notin S, \\ |S| - 1 - \left|\bar{S}\right| - 1 = t - (n-t) - 2 = 2t - n - 2 & \text{otherwise.} \end{cases}$$

If we choose t so that $2t - n = \varepsilon n$ then z is ε-correlated with exactly the points in $\{e_i \mid i \notin S\}$, so it is an $(1 - \frac{t}{n}, \varepsilon)$-center. Since $t/n = (1+\varepsilon)/2$ it is an $(1/2 - \varepsilon/2, \varepsilon)$-center. Note that for z, z' of weight $1/2 + \varepsilon/2$ it holds that whenever $z_i \neq z'_i$ then e_i is either ε-correlated with z or with z' but not with both of them. Therefore, $|C_\varepsilon(z) \triangle C_\varepsilon(z')| = \Delta(z, z')n$ (where $\Delta(z, z')$ is the Hamming distance between z, z').

To get the counterexample,Choose $z^{(1)}, z^{(2)}, \ldots$ to be points that have weight t and are pairwise far from each other as in an error correcting code. ■

Our next result shows that under the promise that the centers are far apart, then the list size is feasible.

Lemma 5.2. *Given a set $X \subset \{-1, 1\}^n$ of N points, then there exist at most t for $t \leq \frac{1}{2\delta\varepsilon}$ centers $z^{(1)}, \ldots, z^{(t)} \in \{-1, 1\}^n$ such that the following holds:*

- *For every $i \in [t]$: $z^{(i)}$ is (ε, δ)-center of X.*
- *For every z that is a (ε, δ)-center of X, there is some i such that $\left\langle z, z^{(i)} \right\rangle > O(\varepsilon^2)$.*

Proof. Assume there is a list of t such centers, we show $t < 1/2\delta\varepsilon$. Using Johnson bound, for each x the number of i's such that $\left\langle x, z^{(i)} \right\rangle > \varepsilon$ is bounded by $1/2\varepsilon$. Now consider a bipartite graph $G = (L, R)$ such that $L = X$ and $R = [t]$, and $(x, i) \in E$ iff $\left\langle x, z^{(i)} \right\rangle > \varepsilon$.

For every $x \in X$, x participates in at most $1/2\varepsilon$ edges. Therefore, the number of edges is bound by $|X|/2\varepsilon$. On the other hand, each $i \in [m]$ has at least $\delta |X|$ edges, so the number of edges is at least $t\delta |X|$. Combining the two inequalities we get $t \leq \frac{1}{2\delta\varepsilon}$. ■

Acknowledgement

We would like to thank Sariel Har-Peled, Guy Kindler and Igor Shinkar for helpful discussions.

References

[AB09] Sanjeev Arora and Boaz Barak, *Computational complexity - a modern approach*, Cambridge University Press, 2009.

[BHPI02] Mihai Badoiu, Sariel Har-Peled, and Piotr Indyk, *Approximate clustering via core-sets*, In Proc. 34th Annu. ACM Sympos. Theory Comput, 2002, pp. 250–257.

[FGKP09] Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami, *On agnostic learning of parities, monomials, and halfspaces*, SIAM J. Comput. **39** (2009), no. 2, 606–645.

[GL89] O. Goldreich and L. A. Levin, *A hard-core predicate for all one-way functions*, Proceedings of the twenty-first annual ACM symposium on Theory of computing (New York, NY, USA), STOC '89, ACM, 1989, pp. 25–32.

[GR06] Venkatesan Guruswami and Prasad Raghavendra, *Hardness of learning halfspaces with noise*, In Proceedings of FOCS, 2006, pp. 543–552.

[IJK06] Russell Impagliazzo, Ragesh Jaiswal, and Valentine Kabanets, *Approximately list-decoding direct product codes and uniform hardness amplification*, Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (Washington, DC, USA), IEEE Computer Society, 2006, pp. 187–196.

[IJKW10] Russell Impagliazzo, Ragesh Jaiswal, Valentine Kabanets, and Avi Wigderson, *Uniform direct product theorems: Simplified, optimized, and derandomized*, vol. 39, 2010, pp. 1637–1665.

[Joh62] S. M. Johnson, *A new upper bound for error-correcting codes*, IRE Transactions on Information Theory **8** (1962), no. 2, 203–207.

[TSSW96] L. Trevisan, G.B. Sorkin, M. Sudan, and D.P. Williamson, *Gadgets approximation, and linear programming*, Foundations of Computer Science, IEEE Annual Symposium on **0** (1996), 617.

A A Detailed Description of [BHPI02]'s Approach

The problem of clustering with outliers was studied in [BHPI02]. They apply the following two steps paradigm: First they show that a random sample of a small subset of points, which they call "coreset", represents the original set well (where by small they mean that the size depends on the accuracy of the approximation and not on the dimension n). Second, they solve the problem on the small set, and show that this solution is a good enough approximation for the original set of points. Let us sketch briefly their approach:

They first show that with constant probability, a random set $S \subset X$ of cardinality $O(1/\delta'\varepsilon')$, satisfies the following:

- There exists $s \in S$, such that there exists a $(\delta - \delta', \varepsilon - \varepsilon')$-center for X that is relatively close to s.
- The linear subspace H spanned by S contains a $(\delta - \delta', \varepsilon - \varepsilon')$-center for X.

They use these two facts in order to build an exponential grid on H. They show that one of the points in the grid is indeed an approximate solution for X.

This approach strongly uses the fact that the points and the centers lie in $\mathbb{R}^n$, and not in the hypercube. In particular [BHPI02] claim that an approximate solution lies in the subspace spanned by S. Even if the original set of points lies in the hypercube, the subspace spanned by them does not lie in the hypercube.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de