Hardness of K-Center Approximation

Sai Praneeth Reddy
2011CS50294

May 1, 2015

Abstract

This is a short report on [1]. We show that approximating intercenter K-means clustering i.e. the problem of minimizing the maximum euclidean intracluster distances is NP-hard for approximation factors $\sqrt{2}-\epsilon$ for all $\epsilon > 0$ when the points belong to $\mathbf{R}^d$ where $d \geq 2$.

1 Overview

The familiar k-center problem in euclidean spaces is changed in two ways in this paper: (i) We replace the notion of distance from a center with the maximum of the distance from all points and (ii) we define it for a general metric space using graphs.

Let $G = (V, E, W)$ be weighted undirected graph with vertex set V, edge set E and distances on the edges (also called a dissimilarity function) W where edge e has distance w_e. A partition of V into clusters $B_1, B_2, \ldots, B_k$ is called a $k - split$. The cost of B_i is the max of $w_e, e \in B_i$, and the cost of a split is the maximum of the costs across all the clusters. The problem is to find a $k - split$ such that the cost is minimized.

We will prove that the above problem is hard to approximate in three steps:

1. We will first define a restricted version of the exact cover problem by 3 sets and show that it is NP-hard.

2. Then given an instance of the restricted cover by 3 sets problem, we convert into a graph with weights which obey triangle inequality and show that finding $(2-\epsilon)$-approximation would lead to solving the problem.

3. We then modify the graph and show how it can be massaged to give us a set of points on a 2D plane whose K-center solution with any approximation better than $\sqrt{2}$ would also lead to solving the restricted cover by 3 size sets problem.

2 The Proof of Hardness

2.1 Restricted Set Cover

In the *exact cover by 3 sets* (also called **XC3**), we are given a set of points $U = \{u_1, \ldots, u_n\}$ and a family of sets defined over these points $\mathbf{F} = \{C_1, \ldots, C_m\}$

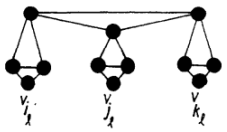

Figure 1: Component representing a triplet C_i in $\mathbf{F}$

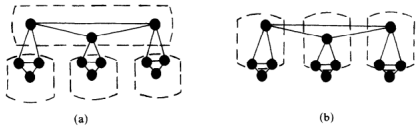

Figure 2: Clusterings of components in (a) $\mathbf{C}$ and in $\mathbf{F} - \mathbf{C}$

where $C_i \subseteq U$ and $|C_i| = 3$. Further it is given that each u_i is present in *atmost* 3 sets in $\mathbf{F}$[1]. The problem is then to find a subset $\mathbf{C} \subseteq \mathbf{F}$ which *exactly* covers U i.e. each u_i is present in exactly one set $C_i \in \mathbf{C}$.

The *exact cover by 3 sets* problem is known to be NP-complete. We will further impose a restriction that each element u_i appears in *exactly* 3 sets in $\mathbf{F}$ called the the *restricted exact cover by 3 sets* (**RXC3**). [2] The problem remains NP-complete. To show this, we present an algorithm *Build* which converts an instance of $XC-3$ to an instance of $RXC-3$. It is straightforward to see that solving $RXC-3$ gives us a solution for $XC-3$.

Algorithm 1 `Build`$(U, \mathtt{F})$

if there is u_i such that it is not present in any set in $\mathbf{F}$ **then**
  return an coverable instance
end if
if The number of elements of U which are present either once or twice in $\mathbf{F}$ is not a multiple of 3 **then**
  Make 3 copies of $(U, \mathbf{F})$ and let $(U, \mathbf{F})$ be the new copy.
end if
while There exists u_i, u_j, u_k which do not appear in exactly 3 subsets in $\mathbf{F}$ **do**
  Add new elements y_i, y_j, y_k to U.
  Add (u_i, y_j, y_k),(u_j, y_k, y_i), and (u_k, y_i, y_j) to $\mathbf{F}$.
end while

2.2 A hard graph

Given an instance of *RXC-3* $(U, \mathbf{F})$, we construct a complete graph $G(V, E, W)$ where w_e for all edges is either 1 or 2.
Vertex Set: One v_i for u_i in U and nine vertices for each set $C_i \in \mathbf{F}$.
Edge Set: It is a complete graph.
Distance weights: For each 3-element C_i, eighteen edges will have weight 1, and the rest will have 2. The edges with weights 1 are shown in figure 1.
k: $k = \frac{9n+n}{3} = \frac{10n}{3}$.

Theorem 1. *A k-split of G with cost 1 is possible* iff *there is an exact cover of* $(u, \mathbf{F})$.

[1] Note that this means $m \leq n$
[2] Now $m = n$.

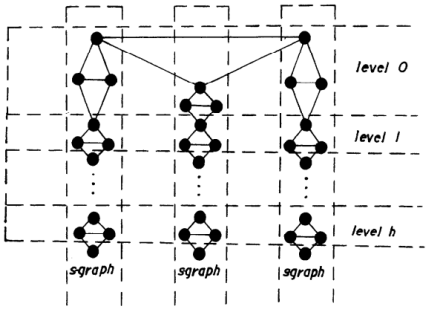

Figure 3: S-graph representing a triplet C_i in $\mathbf{F}$

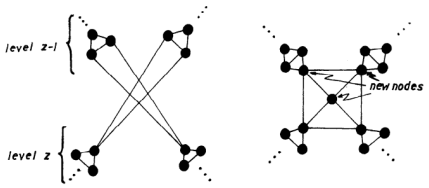

Figure 4: A crossing of an s-graph and its elimination by inserting new nodes

Proof. Given an exact cover solution $\mathbf{C}$, clustering is done according to Figure 2. Then each cluster is a triangle. Thus the cost is . Alternatively, if the cost if 1, then each cluster has to be a triangle as there is no 4-clique. Thus one of the two clusterings in Figure 2 are the only possibilities. As each v_i is in some cluster of type (a) in Fig. 2, pick the corresponding C_i. This forms an exact cover. □

Because all the weights are either 1 or 2, they trivially satisfy the triangle inequality condition. Moreover, the cost of any k-split is either 1 or 2 and so if we can find a solution with approximation factor $(2-\epsilon)$, we would find the optimal solution with cost 1. Hence this construction already gives us that approximating k-center is NP-hard for arbitary metric spaces. The graph can further be embedded in $\mathbf{R}^n$ euclidean space with standard embedding techniques.

2.3 S-graphs

To embed our graph on a 2D plane, we first convert the graph obtained by rules in Fig. 1 to a planar graph. For this, we extend the triplet component to a height $h = O(n^2)$ as shown in Fig 3. This is done so that each of the extension (called an s-graph) intersects another s-graph in exactly one level (see Fig 4 (a)). Such an intersection is taken care by inserting 3 new nodes and changing the edges as shown in Fig 4.

Theorem 2. *A k-split of G' constructed with the new rules with cost 1 is possible* iff *there is an exact cover of* $(u, \mathbf{F})$.

Proof. The reduction occurs similar to the previous case The only problem occurs when the s-graphs intersect each other. As long as the the intersection behaves *properly* i.e. is one of the four types in Fig. 5, the proof of the previous theorem carries through. Given an exact cover $\mathbf{C}$, we cluster into triangles as before. The intersections between the s-graphs are taken care as shown in Fig 5.

Suppose there is a clustering of G' with cost 1. Note that there is still no clique of size 4. Thus all clusters remain triangles. The *bad* cases occur as shown in Fig. 6 (and their mirrored images). However if we cluster them in either manner, there exist three nodes (*exposed nodes*) not all of which can be a part of a triangle, thus making it impossible for the cost to be 1. Hence only

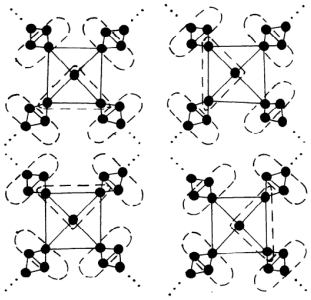

Figure 5: Clustering at the S-graph intersections

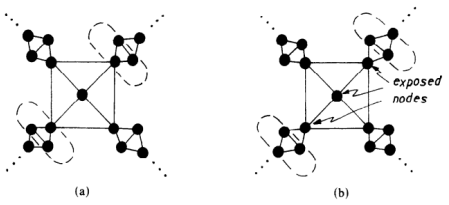

Figure 6: *Bad* clustering are not possible as at least one of the *exposed* nodes will be left out of a cluster

clusters of type shown in Fig. 5 are possible. Using similar arguments as in Theorem 1, we can obtain an exact cover from the clustering. □

2.4 A planar embedding

We now go from the planar graph to embedding points on a 2D plane. Here we will lose the approximation factor we can prove and show that it is hard for any $(\sqrt{2}-\epsilon)$. The graph G' obtained from the *EXC-3* instance is now made up of 4 kinds of blocks. The nodes in each of the blocks are placed as shown in Fig. 7. Note that the lines in the lanar embedding are only indicative of distances and there is no graph (other than the euclidean distance graph implicit in any set of points in $\mathbf{R}^d$).

Theorem 3. *A k-center clustering of the obtained points with cost 1 is possible* iff *there is an exact cover of* $(U, \mathbf{F})$.

Proof. Note that the cost of any triangle is still 1 and there are no 4 nodes whose cluster can have cost 1. After this, the proof is exactly same as the one in Theorem 2. □

This leads us to our final theorem on the hardness of approximation of the k-center problem.

Theorem 4. *Getting a k-center solution is NP-hard for approximation factor* $\sqrt{2}-\epsilon$ *for any* $\epsilon > 0$.

Proof. Note that the implicit complete graph of the point set we defined has edges with weights other 1 or 2. The optimal solution is still ≤ 1 from Theorem 3. The shortest distance between any two points not part of a triangle occurs in the right angled arms in the bottom two transformations (Fig. 7). Their distance using triangle inequality can be seen to be $< \sqrt{2} - \frac{\epsilon}{2} + \frac{\epsilon}{2} = \sqrt{2} - \epsilon$. Thus a solution which gives a solution within an approximation factor of $\sqrt{2}-\epsilon$ would give the optimal solution wit cost 1, and so with Theorem 3, we solve the exact cover problem. □

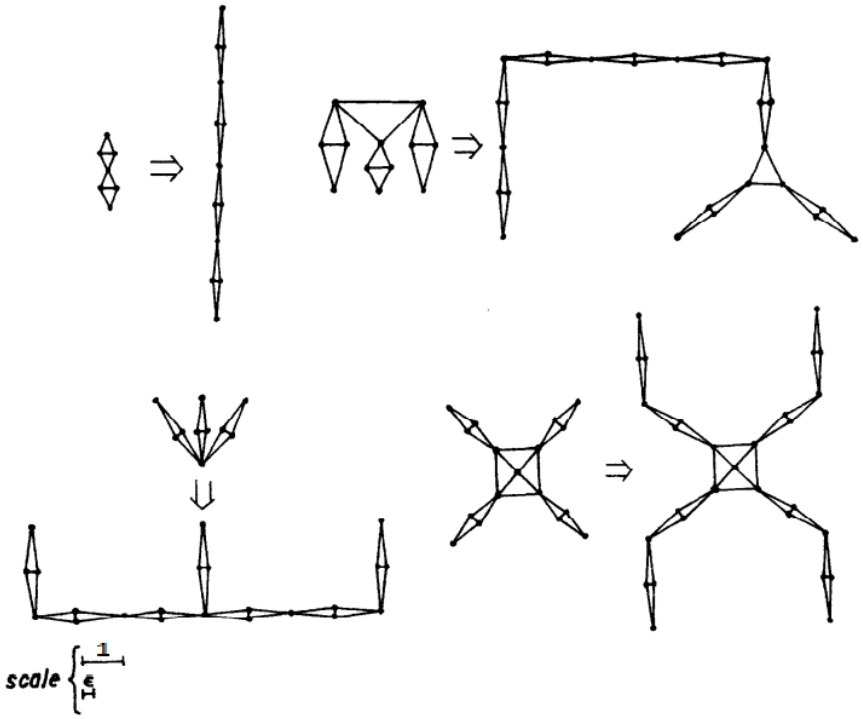

Figure 7: Final transformations to give a set of points embedded on a plane.

References

[1] Gonzalez, Teofilo F, *Clustering to minimize the maximum intercluster distance*, Theoretical Computer Science (Vol. 38) - 1985, pp. 293