Information Processing Letters 120 (2017) 40–43

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

Improved and simplified inapproximability for k-means

Euiwoong Lee [1], Melanie Schmidt [*,2], John Wright [3]

Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213, United States

ARTICLE INFO

Article history:
Received 3 September 2015
Accepted 27 November 2016
Available online 8 December 2016
Communicated by R. Uehara

Keywords:
k-Means
Hardness of approximation
Clustering
Computational complexity

ABSTRACT

The k-means problem consists of finding k centers in $\mathbb{R}^d$ that minimize the sum of the squared distances of all points in an input set P from $\mathbb{R}^d$ to their closest respective center. Awasthi et al. recently showed that there exists a constant $\varepsilon' > 0$ such that it is NP-hard to approximate the k-means objective within a factor of $1 + \varepsilon'$. We establish that $1 + \varepsilon'$ is at least 1.0013.

For a given set of points $P \subset \mathbb{R}^d$, the *k-means problem* consists of finding a partition of P into k clusters $(C_1, \ldots, C_k)$ with corresponding centers $(c_1, \ldots, c_k)$ that minimize the sum of the squared distances of all points in P to their corresponding center, i.e. the quantity

$$\arg \min_{(C_1,\ldots,C_k),(c_1,\ldots,c_k)} \sum_{i=1}^{k} \sum_{x \in C_i} ||x - c_i||^2$$

where $|| \cdot ||$ denotes the Euclidean distance. The k-means problem has been well-known since the fifties, when Lloyd [10] developed the famous local search heuristic also known as the k-means algorithm. Various exact, approximate, and heuristic algorithms have been developed since then. For a constant number of clusters k and a constant dimension d, the problem can be solved by enumerating weighted Voronoi diagrams [7]. If the dimension is arbitrary but the number of centers is constant, many polynomial-time approximation schemes are known. For example, [6] gives an algorithm with running time $\mathcal{O}(nd + 2^{\mathrm{poly}(1/\varepsilon,k)})$. In the general case, only constant-factor approximation algorithms are known [8,9], but no algorithm with an approximation ratio smaller than 9 has yet been found.

Surprisingly, no hardness results for the k-means problem were known even as recently as ten years ago. Today, it is known that the k-means problem is NP-hard, even for constant k and arbitrary dimension d [1,4] and also for arbitrary k and constant d [12]. Early this year, Awasthi et al. [2] showed that there exists a constant $\varepsilon' > 0$ such that it is NP-hard to approximate the k-means objective within a factor of $1 + \varepsilon'$. They use a reduction from the Vertex Cover problem on triangle-free graphs. Here, one is given a graph $G = (V, E)$ that does not contain a triangle, and the goal is to compute a minimal set of vertices S which *covers* all the edges, meaning that for any $(v_i, v_j) \in E$, it holds that $v_i \in S$ or $v_j \in S$. To decide if k vertices suffice to cover a given G, they construct a k-means instance in the following way. Let $b_i = (0, \ldots, 1, \ldots, 0)$ be the ith vector in the standard basis of $\mathbb{R}^{|V|}$. For an edge $e = (v_i, v_j) \in E$, set $x_e = b_i + b_j$. The instance consists of the parameter k and the point set $\{x_e \mid e \in E\}$. Note that the number of points is $|E|$ and their dimension is $|V|$.

* Corresponding author.
E-mail address: melanie.schmidt@tu-dortmund.de (M. Schmidt).

1 Supported by the Samsung Scholarship and NSF CCF-1115525.
2 Supported by the German Academic Exchange Service (DAAD).
3 Supported by a Simons Award for Graduate Students in Theoretical Computer Science.

http://dx.doi.org/10.1016/j.ipl.2016.11.009
0020-0190/

A relatively simple analysis shows that this reduction is approximation-preserving. A vertex cover $S \subseteq V$ of size k corresponds to a solution for k-means where we have centers at $\{b_i : v_i \in S\}$ and each point $x_{(v_i,v_j)}$ is assigned to a center in $S \cap \{b_i, b_j\}$ (which is nonempty because S is a vertex cover). In addition, it can also be shown that a good solution for k-means reveals a small vertex cover of G when G is triangle-free.

Unfortunately, this reduction transforms $(1+\varepsilon)$-hardness for Vertex Cover on triangle-free graphs to $(1+\varepsilon')$-hardness for k-means where $\varepsilon' = O(\frac{\varepsilon}{\Delta})$ and Δ is the maximum degree of G. Awasthi et al. [2] proved hardness of Vertex Cover on triangle-free graphs via a reduction from general Vertex Cover, where the best hardness result of Dinur and Safra [5] has an unspecified large constant Δ. Furthermore, the reduction uses a sophisticated spectral analysis to bound the size of the minimum vertex cover of a suitably chosen graph product.

Our result is based on the observation that hardness results for Vertex Cover on small-degree graphs lead to hardness of Vertex Cover on triangle-free graphs with the same degree in an extremely simple way. Combined with the result of Chlebík and Chlebíková [3] that proves hardness of approximating Vertex Cover on 4-regular graphs within ≈ 1.02, this observation gives hardness of Vertex Cover on triangle-free, degree-4 graphs without relying on the spectral analysis. The same reduction from Vertex Cover on triangle-free graphs to k-means then proves APX-hardness of k-means, with an improved ratio due to the small degree of G.

1. Main result

Our main result is the following theorem.

Theorem 1. *It is NP-hard to approximate k-means within a factor* 1.0013.

We prove hardness of k-means by a reduction from Vertex Cover on 4-regular graphs, for which we have the following hardness result of Chlebík and Chlebíková [3].

Theorem 2 *([3], see also Appendix A). Given a 4-regular graph $G = (V(G), E(G))$, it is NP-hard to distinguish the following cases.*

- *G has a vertex cover with at most $\alpha_{min}|V(G)|$ vertices.*
- *Every vertex cover of G has at least $\alpha_{max}|V(G)|$ vertices.*

Here, $\alpha_{min} = (2\mu_{4,k}+8)/(4\mu_{4,k}+12)$ and $\alpha_{max} = (2\mu_{4,k}+9)/(4\mu_{4,k}+12)$ with $\mu_{4,k} \le 21.7$. In particular, it is NP-hard to approximate Vertex Cover on degree-4 graphs within a factor of $(\alpha_{max}/\alpha_{min}) \ge 1.0192$.

Given a 4-regular graph $G = (V(G), E(G))$ for Vertex Cover with $n := |V(G)|$ vertices and $2n$ edges, we first partition $E(G)$ into E_1 and E_2 such that $|E_1| = |E_2| = |E(G)|/2 = n$ and such that the subgraph $(V(G), E_2)$ is bipartite. Such a partition always exists: every graph has a cut containing at least half of the edges (well-known; see, e.g., [13]). Choose n of these cut edges for E_2 and let E_1 be the remaining edges. We define $G' = (V(G'), E(G'))$ by *splitting* each edge in E_1 into three edges. Formally, G' is given by

$$V(G') = V(G) \cup \left(\bigcup_{e=(u,v)\in E_1} \{v'_{e,u}, v'_{e,v}\} \right),$$

$$E(G') = \left(\bigcup_{e=(u,v)\in E_1} \{(v, v'_{e,v}), (v'_{e,v}, v'_{e,u}), (v'_{e,u}, u)\} \right) \cup E_2.$$

Notice that V has $n + 2n = 3n$ vertices and $3n + n = 4n$ edges. It is also easy to see that the maximum degree of V is 4, and that V does not have any triangle, since any triangle of G contains at least one edge of E_1 (because $(V(G), E_2)$ is bipartite) and each edge of E_1 is split into three.

Given G' as an instance of Vertex Cover on triangle-free graphs, the reduction to the k-means problem is the same as before. Let $b_i = (0, \dots, 1, \dots, 0)$ be the ith vector in the standard basis of $\mathbb{R}^{3n}$. For an edge $e = (v_i, v_j) \in E(G')$, set $x_e = b_i + b_j$. The instance consists of the parameter $k = (\alpha_{min}+1)n$ and the point set $\{x_e \mid e \in E\}$. Notice that the number of points is now $4n$ and their dimension is $3n$.

We now analyze the reduction. Note that for k-means, once a cluster is fixed as a set of points, the optimal center and the cost of the cluster are determined.[4] Let $\mathrm{cost}(C)$ be the cost of a cluster C. We abuse notation and use C for the set of edges $\{e : x_e \in C\} \subseteq E(G')$ as well. For an integer l, define an *l-star* to be a set of l distinct edges incident to a common vertex. The following lemma is proven by Awasthi et al. and shows that if C is cost-efficient, then two vertices are sufficient to cover many edges in C. Furthermore, an *optimal* C is either a star or a triangle.

Lemma 3 *([2], Proposition 9 and Lemma 11). Let $C = \{x_{e_1}, \dots, x_{e_l}\}$ be a cluster. Then $l - 1 \le \mathrm{cost}(C) \le 2l - 1$, and there exist two vertices that cover at least $\lceil 2l - 1 - \mathrm{cost}(C) \rceil$ edges in C. Furthermore, $\mathrm{cost}(C) = l - 1$ if and only if C is either an l-star or a triangle, and otherwise, $\mathrm{cost}(C) \ge l - 1/2$.*

1.1. Completeness

Lemma 4. *If G has a vertex cover of size at most $\alpha_{min}n$, the instance of k-means produced by the reduction admits a solution of cost at most $(3 - \alpha_{min})n$.*

Proof. Suppose G has a vertex cover S with at most $\alpha_{min}n$ vertices. For each edge $e = (u, v) \in E_1$, let $v'(e) = v'_{e,u}$ if $v \in S$, and $v'(e) = v'_{e,v}$ otherwise. Let $S' := S \cup (\cup_{e \in E_1}\{v'(e)\})$. Since S is a vertex cover of G, for every edge $e \in E_1$, S and $v'(e)$ cover all three edges of $E(G')$ corresponding to e. Therefore, S' is a vertex cover of G', and since $|E_1| = n$, it has at most $(\alpha_{min}+1)n$ vertices.

[4] For $k = 1$, the optimal solution to the k-means problem is the *centroid* of the point set. This is due to a well-known fact, see, e.g., Lemma 2.1 in [9].