Information Processing Letters 176 (2022) 106251

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

A refined approximation for Euclidean k-means

Fabrizio Grandoni [a,*], Rafail Ostrovsky [b], Yuval Rabani [c], Leonard J. Schulman [d], Rakesh Venkat [e]

[a] IDSIA, USI-SUPSI, Switzerland
[b] UCLA, United States of America
[c] The Hebrew University of Jerusalem, Israel
[d] Caltech, United States of America
[e] IIT Hyderabad, India

ARTICLE INFO

Article history:
Received 15 July 2021
Accepted 23 January 2022
Available online 2 February 2022
Communicated by Leah Epstein

Keywords:
Approximation algorithms
Euclidean k-means
Euclidean facility location
Integrality gaps

ABSTRACT

In the Euclidean k-Means problem we are given a collection of n points $\mathcal{D}$ in an Euclidean space and a positive integer k. Our goal is to identify a collection of k points in the same space (centers) so as to minimize the sum of the squared Euclidean distances between each point in $\mathcal{D}$ and the closest center. This problem is known to be APX-hard and the current best approximation ratio is a primal-dual 6.357 approximation based on a standard LP for the problem [Ahmadian et al. FOCS'17, SICOMP'20].

In this note we show how a minor modification of Ahmadian et al.'s analysis leads to a slightly improved 6.12903 approximation. As a related result, we also show that the mentioned LP has integrality gap at least $\frac{16+\sqrt{5}}{15} > 1.2157$.

1. Introduction

Clustering is a central problem in Computer Science, with many applications in data science, machine learning etc. One of the most famous and best-studied problems in this area is Euclidean k-Means: given a set $\mathcal{D}$ of n points (or *demands*) in $\mathbb{R}^\ell$ and an integer $k \in [1, n]$, select k points S (*centers*) so as to minimize $\sum_{j\in\mathcal{D}} d^2(j, S)$. Here $d(j, i)$ is the Euclidean distance between points j and i and for a set of points I, $d(j, I) = \min_{i\in I} d(j, i)$. In other words, we wish to select k centers so as to minimize the sum of the squared Euclidean distances between each demand and the closest center. Equivalently, a feasible solution is given by a partition of the demands into k subsets (*clusters*). The cost w_C of a cluster $C \subset \mathcal{D}$ is $\sum_{j\in C} d^2(c, \mu)$, where μ is the center of mass of C. We recall that w_C can also be expressed as $\frac{1}{2|C|}\sum_{j\in C}\sum_{j'\in C} d^2(j, j')$. Our goal is to minimize the total cost of these clusters.

Euclidean k-Means is well-studied in terms of approximation algorithms. It is known to be APX-hard. More precisely, it is hard to approximate k-Means below a factor 1.0013 in polynomial time unless $P = NP$ [6,16]. The hardness was improved to 1.07 under the Unique Games Conjecture [9]. Some heuristics are known to perform very well in practice, however their approximation factor is $O(\log k)$ or worse on general instances [3,4,17,21]. Constant approximation algorithms are known. A local-search algorithm by Kanugo et al. [15] provides a $9 + \varepsilon$ approximation.[1] The authors also show that natural local-search based algorithms cannot perform better than this. This ratio was improved to 6.357 by Ahmadian et al. [1,2]

* Corresponding author.
E-mail address: fabrizio@idsia.ch (F. Grandoni).

[1] Throughout this paper by ε we mean an arbitrarily small positive constant. W.l.o.g. we assume $\varepsilon \leq 1$.

https://doi.org/10.1016/j.ipl.2022.106251

using a primal-dual approach. They also prove a $9+\varepsilon$ approximation for general (possibly non-Euclidean) metrics. Better approximation factors are known under reasonable restrictions on the input [5,7,10,20]. A PTAS is known for constant k [19] or for constant dimension ℓ [10,12]. Notice that ℓ can be always assumed to be $O(\log n)$ by a standard application of the Johnson-Lindenstrauss transform [14]. This was recently improved to $O(\log k + \log\log n)$ [8] and finally to $O(\log k)$ [18].

In this paper we describe a simple modification of the analysis of Ahmadian et al. [2] which leads to a slightly improved approximation for Euclidean k-Means (see Section 2).

Theorem 1. *There exists a deterministic polynomial-time algorithm for Euclidean k-Means with approximation ratio $\rho+\varepsilon$ for any positive constant $\varepsilon > 0$, where*

$$\rho := \left(1+\sqrt{\frac{1}{2}(2+\sqrt[3]{3-2\sqrt{2}}+\sqrt[3]{3+2\sqrt{2}})}\right)^2 < 6.12903.$$

The above approximation ratio is w.r.t. the optimal fractional solution to a standard LP relaxation $LP_{\text{k-Means}}$ for the problem (defined later). As a side result (see Section 3), we prove a lower bound on the integrality gap of this relaxation (we are not aware of any explicit such lower bound in the literature).

Theorem 2. *The integrality gap of $LP_{k\text{-Means}}$, even in the Euclidean plane (i.e., for $\ell = 2$), is at least $\frac{16+\sqrt{5}}{15} > 1.2157$.*

1.1. Preliminaries

As mentioned earlier, one can formulate Euclidean k-Means in term of the selection of k centers. In this case, it is convenient to discretize the possible choices for the centers, hence obtaining a polynomial-size set $\mathcal{F}$ of candidate centers, at the cost of an extra factor $1+\varepsilon$ in the approximation ratio (we will neglect this factor in the approximation ratios since it is absorbed by analogous factors in the rest of the analysis). In particular we will use the construction in [11] (Lemma 24) that chooses as $\mathcal{F}$ the centers of mass of any collection of up to $16/\varepsilon^2$ points with repetitions. In particular $|\mathcal{F}| = O(n^{16/\varepsilon^2})$ in this case.

Let $c(j,i)$ be an abbreviation for $d^2(j,i)$. Then a standard LP-relaxation for k-Means is as follows:

$$\begin{aligned}
&\min \sum_{i\in\mathcal{F}, j\in\mathcal{D}} x_{ij}\cdot c(j,i) && LP_{\text{k-Means}}\\
&s.t. \sum_{i\in\mathcal{F}} x_{ij} \geq 1 && \forall j\in\mathcal{D}\\
&x_{ij} \leq y_i && \forall j\in\mathcal{D}, \forall i\in\mathcal{F}\\
&\sum_{i\in\mathcal{F}} y_i \leq k && \forall j\in\mathcal{D}, \forall i\in\mathcal{F}\\
&x_{ij}, y_i \geq 0 && \forall j\in\mathcal{D}, \forall i\in\mathcal{F}
\end{aligned}$$

In an integral solution, we interpret $y_i = 1$ as i being a selected center in S (i is *open*), and $x_{ij} = 1$ as demand j being *assigned* to center i.[2] The first family of constraints states that each demand has to be assigned to some center, the second one that a demand can only be assigned to an open center, and the third one that we can open at most k centers.

For any parameter $\lambda > 0$ (*Lagrangian multiplier*), the *Lagrangian relaxation* $LP(\lambda)$ of $LP_{\text{k-Means}}$ (w.r.t. the last matrix constraint) and its dual $DP(\lambda)$ are as follows:

$$\begin{aligned}
&\min \sum_{i\in\mathcal{F}, j\in\mathcal{D}} x_{ij}\cdot c(j,i) + \lambda\cdot\sum_{i\in\mathcal{F}} y_i - \lambda k && LP(\lambda)\\
&s.t. \sum_{i\in\mathcal{F}} x_{ij} \geq 1 && \forall j\in\mathcal{D}\\
&x_{ij} \leq y_i && \forall j\in\mathcal{D}, \forall i\in\mathcal{F}\\
&x_{ij}, y_i \geq 0 && \forall j\in\mathcal{D}, \forall i\in\mathcal{F}
\end{aligned}$$

[2] Technically each demand is automatically assigned to the closest open center. However it is convenient to allow also sub-optimal assignments in the LP relaxation.

$$\begin{aligned} &\max \sum_{j\in\mathcal{D}} \alpha_j - \lambda k && DP(\lambda) \\ &s.t. \sum_{j\in\mathcal{D}} \max\{0, \alpha_j - c(j,i)\} \le \lambda && \forall i \in \mathcal{F} \\ &\alpha_j \ge 0 && \forall j \in \mathcal{D} \end{aligned} \tag{1}$$

Above $\max\{0, \alpha_j - c(j,i)\}$ replaces the dual variable β_{ij} corresponding to the second constraint in the primal in the standard formulation of the dual LP. Notice that, by removing the fixed term $-\lambda k$ in the objective functions of $LP(\lambda)$ and $DP(\lambda)$, one obtains the standard LP relaxation $LP_{FL}(\lambda)$ for the Facility Location problem (FL) with uniform facility cost λ and its dual $DP_{FL}(\lambda)$.

We say that a ρ-approximation algorithm for a FL instance of the above type is *Lagrangian Multiplier Preserving* (LMP) if it returns a set of facilities S that satisfies:

$$\sum_{j\in\mathcal{D}} c(j,S) \le \rho(OPT(\lambda) - \lambda|S|),$$

where $OPT(\lambda)$ is the value of the optimal solution to $LP_{FL}(\lambda)$.

2. A refined approximation for Euclidean *k*-means

In this section we present our refined approximation for Euclidean k-Means. We start by presenting the LMP approximation algorithm for the FL instances arising from k-Means described in [2] in Section 2.1. We then present the analysis of that algorithm as in [2] in Section 2.2. In Section 2.3 we describe our refined analysis of the same algorithm. Finally, in Section 2.4 we sketch how to use this to approximate k-Means.

2.1. A primal-dual LMP algorithm for Euclidean facility location

We consider an instance of Euclidean FL induced by a k-Means instance in the mentioned way, for a given Lagrangian multiplier $\lambda > 0$.

We consider exactly the same Lagrangian Multiplier Preserving (LMP) primal-dual algorithm $JV(\delta)$ as in [2]. In more detail, let $\delta \ge 2$ be a parameter to be fixed later. The algorithm consists of a dual-growth phase and a pruning phase. The dual-growth phase is exactly as in the classical primal-dual algorithm JV by Jain and Vazirani [13]. We start with all the dual variables set to 0 and an empty set $\mathcal{O}_t$ of tentatively open facilities. The clients such that $\alpha_j \ge c(j,i)$ for some $i \in \mathcal{O}_t$ are frozen, and the other clients are active. We grow the dual variables of active clients at uniform rate until one of the following two events happens. The first event is that some constraint of type (1) becomes tight. At that point the corresponding facility i is added to $\mathcal{O}_t$ and all clients j with $\alpha_j \ge c(j,i)$ are set to frozen. The second event is that $\alpha_j = c(j,i)$ for some $i \in \mathcal{O}_t$. In that case j is set to frozen. In any case, the facility $w(j)$ that causes j to become frozen is called the witness of j. The phase halts when all clients are frozen.

In the pruning phase we will close some facilities in $\mathcal{O}_t$, hence obtaining the final set of open facilities IS. Here $JV(\delta)$ deviates from JV. For each client $j \in \mathcal{D}$, let $N(j) = \{i \in \mathcal{F} : \alpha_j > c(j,i)\}$ be the set of facilities i such that j contributed with a positive amount to the opening of i. Symmetrically, for $i \in \mathcal{F}$, let $N(i) = \{j \in \mathcal{D} : \alpha_j > c(j,i)\}$ be the clients that contributed with a positive amount to the opening of i. For $i \in \mathcal{O}_t$, we let $t_i = \max_{j\in N(i)} \alpha_j$, where the values α_j are considered at the end of the dual-growth phase. We set conventionally $t_i = 0$ for $N(i) = \emptyset$. Intuitively, t_i is the "time" when facility i is tentatively open (at which point all the dual variables of contributing clients stop growing). We define a conflict graph H over tentatively open facilities as follows. The node set of H is $\mathcal{O}_t$. We place an edge between $i, i' \in \mathcal{O}_t$ iff the following two conditions hold: (1) for some client j, $j \in N(i) \cap N(i')$ (in words, j contributes to the opening of both i and i') and (2) one has $c(i,i') \le \delta \cdot \min\{t_i, t_{i'}\}$. In this graph we compute a maximal independent set IS, which provides the desired solution to the facility location problem (where each client is assigned to the closest facility in IS).

We remark that the pruning phase of JV differs from the one of $JV(\delta)$ only in the definition of H, where condition (2) is not required to hold (or, equivalently, JV behaves like $JV(+\infty)$ for $\lambda > 0$).

2.2. The analysis in [2]

The general goal is to show that

$$\sum_{j\in\mathcal{D}} c(j,IS) \le \rho(\sum_{j\in\mathcal{D}} \alpha_j - \lambda|IS|),$$

for some $\rho \ge 1$ as small as possible. This shows that the algorithm is an LMP ρ-approximation for the problem. It is sufficient to prove that, for each client j, one has

$$\frac{c(j,IS)}{\rho} \le \alpha_j - \sum_{i\in N(j)\cap IS}(\alpha_j - c(j,i)) = \alpha_j - \sum_{i\in IS}\max\{0,\alpha_j - c(j,i)\}.$$

Let $S = N(j)\cap IS$ and $s = |S|$. We distinguish 3 cases depending on the value of s.

Case A: $s = 1$ Let $S = \{i^*\}$. Then for any $\rho \ge 1$,

$$\frac{c(j,IS)}{\rho} \le c(j,IS) = c(j,i^*) = \alpha_j - (\alpha_j - c(j,i^*)).$$

Case B: $s > 1$ Here we use the properties of Euclidean metrics. The sum $\sum_{i\in S} c(j,i)$ is the sum of the squared distances from j to the facilities in S. This quantity is lower bounded by the sum of the squared distances from S to the centroid μ of S. Recall that $\sum_{i\in S} c(\mu,i) = \frac{1}{2s}\sum_{i\in S}\sum_{i'\in S} c(i,i')$. We also observe that, by construction, for any two distinct $i,i'\in IS$ one has

$$c(i,i') > \delta\cdot\min\{t_i,t_{i'}\} \ge \delta\cdot\alpha_j,$$

where the last inequality follows from the fact that j is contributing to the opening of both i and i'. Altogether one obtains

$$\sum_{i\in S} c(j,i) \ge \sum_{i\in S} c(\mu,i) = \frac{1}{2s}\sum_{i\in S}\sum_{i'\in S} c(i,i') \ge \frac{(s-1)\delta\alpha_j}{2}.$$

Thus

$$\sum_{i\in S}(\alpha_j - c(j,i)) \le (s - \frac{\delta(s-1)}{2})\alpha_j = (s(1-\frac{\delta}{2}) + \frac{\delta}{2})\alpha_j \overset{\delta\ge 2, s\ge 2}{\le} (2-\frac{\delta}{2})\alpha_j.$$

Using the fact that $\alpha_j > c(j,i)$ for all $i\in S$, hence $\alpha_j > c(j,IS)$, one gets

$$(\frac{\delta}{2}-1)c(j,IS) \overset{\delta\ge 2}{\le} (\frac{\delta}{2}-1)\alpha_j.$$

We conclude that

$$\sum_{i\in S}(\alpha_j - c(j,i)) + (\frac{\delta}{2}-1)c(j,IS) \le (2-\frac{\delta}{2})\alpha_j + (\frac{\delta}{2}-1)\alpha_j = \alpha_j.$$

This gives the desired inequality assuming that $\rho \ge \frac{1}{\delta/2-1}$.

Case C: $s = 0$ Consider the witness $i = w(j)$ of j. Notice that $\alpha_j \ge t_i$ and $\alpha_j \ge c(j,i) = d^2(j,i)$. Hence

$$d(j,i) + \sqrt{\delta t_i} \le (1+\sqrt{\delta})\sqrt{\alpha_j}.$$

If $i\in IS$, then $d(j,IS) \le d(j,i)$. Otherwise there exists $i'\in IS$ such that $d^2(i,i') \le \delta\min\{t_i,t_{i'}\} \le \delta t_i$. Thus $d(j,IS) \le d(j,i) + d(i,i') \le d(j,i) + \sqrt{\delta t_i}$. In both cases one has $d(j,IS) \le (1+\sqrt{\delta})\sqrt{\alpha_j}$, hence

$$c(j,IS) \le (1+\sqrt{\delta})^2\alpha_j.$$

This gives the desired inequality for $\rho \ge (1+\sqrt{\delta})^2$.

Fixing δ Altogether we can set $\rho = \max\{\frac{1}{\delta/2-1}, (1+\sqrt{\delta})^2\}$. The best choice for δ (namely, the one that minimizes ρ) is the solution of $\frac{1}{\delta/2-1} = (1+\sqrt{\delta})^2$. This is achieved for $\delta \simeq 2.3146$ and gives $\rho \simeq 6.3574$.

2.3. A refined analysis

We refine the analysis in Case B as follows. Let $\Delta = \sum_{i\in S} c(j,i)$. Our goal is to upper bound

$$\frac{c(j,S)}{\alpha_j - \sum_{i\in S}(\alpha_j - c(j,i))} = \frac{c(j,S)}{\sum_{i\in S} c(j,i) - (s-1)\alpha_j} = \frac{c(j,S)}{\Delta - (s-1)\alpha_j}.$$

Instead of using the upper bound $c(j,S) \le \alpha_j$ we use the average

$$c(j,S) \le \frac{1}{s}\sum_{i\in S} c(j,i) = \frac{\Delta}{s}.$$

Then it is sufficient to upper bound

$$\frac{1}{s}\frac{\Delta}{\Delta-(s-1)\alpha_j}.$$

The derivative in Δ of the above function is $\frac{1}{s}\frac{-(s-1)\alpha_j}{(\Delta-(s-1)\alpha_j)^2} < 0$. Hence the maximum is achieved for the smallest possible value of Δ. Recall that we already showed that $\Delta \geq \frac{(s-1)\delta\alpha_j}{2}$. Hence a valid upper bound is

$$\frac{1}{s}\frac{\frac{(s-1)\delta\alpha_j}{2}}{\frac{(s-1)\delta\alpha_j}{2}-(s-1)\alpha_j} = \frac{1}{s}\frac{\delta/2}{\delta/2-1} \overset{s\geq 2,\delta\geq 2}{\leq} \frac{\delta/4}{\delta/2-1}.$$

This imposes $\rho \geq \frac{\delta/4}{\delta/2-1}$ rather than $\rho \geq \frac{1}{\delta/2-1}$ in Case B. Notice that this is an improvement for $\delta < 4$. The best choice of δ is now obtained by imposing $\frac{\delta/4}{\delta/2-1} = (1+\sqrt{\delta})^2$. This gives $\delta = \frac{1}{2}(2+\sqrt[3]{3-2\sqrt{2}}+\sqrt[3]{3+2\sqrt{2}}) \simeq 2.1777$ and $\rho = \left(1+\sqrt{\frac{1}{2}(2+\sqrt[3]{3-2\sqrt{2}}+\sqrt[3]{3+2\sqrt{2}})}\right)^2 < 6.12903$.

2.4. From facility location to k-means

We can use the refined $\rho := \left(1+\sqrt{\frac{1}{2}(2+\sqrt[3]{3-2\sqrt{2}}+\sqrt[3]{3+2\sqrt{2}})}\right)^2$ approximation for Euclidean Facility Location from previous section to derive a $\rho+\varepsilon$ approximation for Euclidean k-Means, for any constant $\varepsilon > 0$. Here we follow the approach of [2] with only minor changes. In more detail, the authors consider a variant of the FL algorithm described before, whose approximation factor is $\rho+\varepsilon$ rather than ρ. A careful use of this algorithm leads to a solution opening precisely k facilities, which leads to the desired approximation factor. In their analysis the authors use slight modifications of the inequality $(1+\sqrt{\delta})\sqrt{\alpha_j} \geq d(j,i)+\sqrt{\delta t_i}$ (coming from Case C, which is the same in their and our analysis). The goal is to prove that the modified algorithm is $\rho+\varepsilon$ approximate. Here δ and ρ are used as parameters. Therefore it is sufficient to replace their values of these parameters with the ones coming from our refined analysis. The rest is identical.

3. Lower bound on the integrality gap

In this section we describe our lower bound instance for the integrality gap of $LP_{\text{k-Means}}$. It is convenient to consider first the following slightly different relaxation, based on clusters (with w_C as defined in Section 1):

$$\begin{aligned}
&\min \sum_{C\in\mathcal{C}} w_C x_C && LP'_{\text{k-Means}}\\
&s.t. \sum_{C\in\mathcal{C}:j\in C} x_C \geq 1 && \forall j\in\mathcal{D}\\
&\sum_{C\in\mathcal{C}} x_C \leq k &&\\
&x_C \geq 0 && \forall C\in\mathcal{C}
\end{aligned}$$

Here $\mathcal{C}$ denotes the set of possible clusters, i.e. the possible subsets of points. In an integral solution $x_C = 1$ means that cluster C is part of our solution.

Our instance is on the Euclidean plane, and its points are the (10) vertices of two regular pentagons of side length 1. These pentagons are placed so that any two vertices of distinct pentagons are at large enough distance M to be fixed later. Here $k = 5$. We remark that our argument can be easily extended to an arbitrary number of points by taking $2h$ such pentagons for any integer $h \geq 1$ so that the pairwise distance between vertices of distinct pentagons is at least M, and setting $k = 5h$.

A feasible fractional solution is obtained by setting $x_C = 0.5$ for every C consisting of a pair of consecutive vertices in the same pentagon (so we are considering 10 fractional clusters in total). Obviously this solution is feasible. The cost w_C of each such cluster C is $2(0.5)^2 = 0.5$. Hence the cost of this fractional solution is $10\cdot 0.5\cdot 0.5 = \frac{5}{2}$.

Next consider the optimal integral solution, consisting of 5 clusters. Recall that the radius of each pentagon (i.e. the distance from a vertex to its center) is $r = \sqrt{\frac{2}{5-\sqrt{5}}} \simeq 0.851$ and the distance between two non-consecutive vertices in the same pentagon is $d = \frac{\sqrt{5}+1}{2} \simeq 1.618$. A solution with two clusters consisting of the vertices of each pentagon costs $10r^2$. Any cluster involving vertices of distinct pentagons costs at least $M^2/2$, hence for M large enough the optimal solution forms clusters only with vertices of the same pentagon. In more detail the optimal solution consists of $x \in \{1,2,3,4\}$ clusters containing the vertices of one pentagon and $5-x$ clusters containing the vertices of the remaining pentagon.

Let $w(x)$ be the minimum cost associated with one pentagon assuming that we form x clusters with its vertices. Clearly $w(1) = 5r^2 = \frac{10}{5-\sqrt{5}}$. Regarding $w(4)$, it is obviously convenient to choose two consecutive vertices in the unique cluster of size 2. Thus $w(4) = 1/2$. For $x \in \{2, 3\}$, we note, as is easy to verify, that clusters with consecutive vertices are less expensive than the alternatives. For $w(2)$, one might form one cluster of size 1 and one of size 4. This would cost $\frac{3(1+d^2)}{4} = \frac{15+3\sqrt{5}}{8}$. Alternatively, one might form one cluster of size 2 and one of size 3, at smaller cost $\frac{1}{2} + \frac{2+d^2}{3} = \frac{10+\sqrt{5}}{6}$. Thus $w(2) = \frac{10+\sqrt{5}}{6}$. For $x = 3$, one might form two clusters of size 1 and one of size 3, or two clusters of size 2 and one of size 1. The associated cost in the two cases is $\frac{2+d^2}{3} > 1$ and $2 \cdot \frac{1}{2} = 1$, resp. Hence $w(3) = 1$. So the overall cost of the optimal integral solution is $\min\{w(1) + w(4), w(2) + w(3)\} = w(2) + w(3) = \frac{16+\sqrt{5}}{6}$. Thus the integrality gap of $LP'_{\text{k-Means}}$ is at least $\frac{16+\sqrt{5}}{6} \cdot \frac{2}{5} = \frac{16+\sqrt{5}}{15}$.

Consider next $LP_{\text{k-Means}}$. Here a technical complication comes from the definition of $\mathcal{F}$ which is not part of the input instance of k-Means. The same construction as above works if we let $\mathcal{F}$ contain the centers of mass of any set of 2 or 3 points. Notice that this is automatically guaranteed by the construction in [11] for $16/\varepsilon^2 \geq 3$. In this case the optimal integral solutions to $LP_{\text{k-Means}}$ and $LP'_{\text{k-Means}}$ are the same in the considered example. Furthermore one obtains a feasible fractional solution to $LP_{\text{k-Means}}$ of cost $5/2$ by setting $y_i = 0.5$ for the centers of mass of any two consecutive vertices of the same pentagon, and setting $x_{ij} = 0.5$ for each point i and the two closest centers j with positive y_j. This concludes the proof of Theorem 2.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Work supported in part by the NSF grants 1909972 and CNS-2001096, the SNF excellence grant 200020B_182865/1, the BSF grants 2015782 and 2018687, and the ISF grants 956-15, 2533-17, and 3565-21.

References

[1] Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, Justin Ward, Better guarantees for k-means and Euclidean k-median by primal-dual algorithms, in: Chris Umans (Ed.), 58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017, IEEE Computer Society, , 2017, pp. 61–72.
[2] Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, Justin Ward, Better guarantees for k-means and Euclidean k-median by primal-dual algorithms, SIAM J. Comput. 49 (4) (2020).
[3] David Arthur, Sergei Vassilvitskii, How slow is the k-means method?, in: Nina Amenta, Otfried Cheong (Eds.), Proceedings of the 22nd ACM Symposium on Computational Geometry, Sedona, Arizona, USA, June 5–7, 2006, ACM, 2006, pp. 144–153.
[4] David Arthur, Sergei Vassilvitskii, k-means++: the advantages of careful seeding, in: Nikhil Bansal, Kirk Pruhs, Clifford Stein (Eds.), Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007, SIAM, 2007, pp. 1027–1035.
[5] Pranjal Awasthi, Avrim Blum, Or Sheffet, Stability yields a PTAS for k-median and k-means clustering, in: 51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA, IEEE Computer Society, 2010, pp. 309–318.
[6] Pranjal Awasthi, Moses Charikar, Ravishankar Krishnaswamy, Ali Kemal Sinop, The hardness of approximation of Euclidean k-means, in: Lars Arge, János Pach (Eds.), 31st International Symposium on Computational Geometry, SoCG 2015, June 22-25, 2015, Eindhoven, The Netherlands, in: LIPIcs, vol. 34, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2015, pp. 754–767.
[7] Maria-Florina Balcan, Avrim Blum, Anupam Gupta, Approximate clustering without the approximation, in: Claire Mathieu (Ed.), Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2009, New York, NY, USA, January 4-6, 2009, SIAM, 2009, pp. 1068–1077.
[8] Luca Becchetti, Marc Bury, Vincent Cohen-Addad, Fabrizio Grandoni, Chris Schwiegelshohn, Oblivious dimension reduction for k-means: beyond subspaces and the Johnson-Lindenstrauss lemma, in: Moses Charikar, Edith Cohen (Eds.), Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019, ACM, 2019, pp. 1039–1050.
[9] Vincent Cohen-Addad, C.S. Karthik, Inapproximability of clustering in l_p metrics, in: David Zuckerman (Ed.), 60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019, Baltimore, Maryland, USA, November 9-12, 2019, IEEE Computer Society, 2019, pp. 519–539.
[10] Vincent Cohen-Addad, Philip N. Klein, Claire Mathieu, Local search yields approximation schemes for k-means and k-median in Euclidean and minor-free metrics, SIAM J. Comput. 48 (2) (2019) 644–667.
[11] Wenceslas Fernandez de la Vega, Marek Karpinski, Claire Kenyon, Yuval Rabani, Approximation schemes for clustering problems, in: Lawrence L. Larmore, Michel X. Goemans (Eds.), Proceedings of the 35th Annual ACM Symposium on Theory of Computing, June 9-11, 2003, San Diego, CA, USA, ACM, 2003, pp. 50–58.
[12] Zachary Friggstad, Mohsen Rezapour, Mohammad R. Salavatipour, Local search yields a PTAS for k-means in doubling metrics, SIAM J. Comput. 48 (2) (2019) 452–480.
[13] Kamal Jain, Vijay V. Vazirani, Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and Lagrangian relaxation, J. ACM 48 (2) (2001) 274–296.
[14] William B. Johnson, Joram Lindenstrauss, Extensions of Lipschitz mappings into a Hilbert space, Contemp. Math. 26 (1984) 189–206.
[15] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, Angela Y. Wu, A local search approximation algorithm for k-means clustering, Comput. Geom. 28 (2–3) (2004) 89–112.
[16] Euiwoong Lee, Melanie Schmidt, John Wright, Improved and simplified inapproximability for k-means, Inf. Process. Lett. 120 (2017) 40–43.
[17] Stuart P. Lloyd, Least squares quantization in PCM, IEEE Trans. Inf. Theory 28 (2) (1982) 129–136.
[18] Konstantin Makarychev, Yury Makarychev, Ilya P. Razenshteyn, Performance of Johnson-Lindenstrauss transform for k-means and k-medians clustering, in: Moses Charikar, Edith Cohen (Eds.), Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019, ACM, 2019, pp. 1027–1038.

[19] Jirí Matousek, On approximate geometric k-clustering, Discrete Comput. Geom. 24 (1) (2000) 61–84.
[20] Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, Chaitanya Swamy, The effectiveness of Lloyd-type methods for the k-means problem, J. ACM 59 (6) (2012) 28.
[21] Andrea Vattani, k-means requires exponentially many iterations even in the plane, Discrete Comput. Geom. 45 (4) (2011) 596–616.