The hardness of k-means clustering in the plane

Andrea Vattani
University of California, San Diego
avattani@ucsd.edu

Abstract

We show that k-means clustering is an NP-hard optimization problem, even for instances in the plane. Specifically, the hardness holds for $k = \Theta(n^{\epsilon})$, for any $\epsilon > 0$, where n is the number of points in the instance, and k is the number of clusters.

1 Introduction

In the k-means clustering problem we are given a finite set of points S in $\mathbb{R}^d$, an integer $k \geq 1$, and the goal is to find k points (usually called *centers*) so to minimize the sum of the squared Euclidean distance of each point in S to its closest center. In this brief note, we will show that k-means clustering is NP-hard even in $d = 2$ dimensions. This result has also been shown by Mahajan, Nimbhorkar, and Varadarajan [3] with a reduction from Planar 3-SAT. Our proof instead will use a reduction from Exact Cover by 3-Sets (X3C). We will prove the hardness by considering k-means instances with weighted points. We observe that this is without loss of generality since we can replace a point x of weight w with w distinct points very close to x.

We define the decisional version of the weighted k-means clustering problem.

Definition 1. *Given a multiset $S \subseteq \mathbb{R}^d$, an integer k and $L \in \mathbb{R}$, is there a subset $T \subset \mathbb{R}^d$ with $|T| = k$ such that $\sum_{x \in S} \min_{t \in T} ||x - t||^2 \leq L$?*

Our theorem is the following.

Theorem 2. *The k-means clustering problem is NP-complete even for $d = 2$.*

It is easy to see that the so defined k-means clustering is in NP. To show that it is indeed NP-complete we will reduce from the Exact Cover by 3-Sets problem (X3C) which is known to be NP-complete [1]. It is defined in the following way.

Definition 3. *Given a finite set U containing exactly $3n$ elements and a collection $\mathcal{C} = \{S_1, S_2, \ldots, S_l\}$ of subsets of U each of which contains exactly 3 elements, are there n sets in $\mathcal{C}$ such that their union is U?*

During the analysis, we will make extensive use of the well-known property that, for the k-means cost function, the cost of a cluster C can be computed as $\frac{1}{|C|} \sum_{\{x,y\} \in \binom{C}{2}} ||x - y||^2$ (see for example [2]). For weighted points this translates to

$$\frac{1}{\sum_{x \in C} w(x)} \sum_{\{x,y\} \in \binom{C}{2}} w(x)w(y)||x - y||^2 \tag{1}$$

where $w(x)$ denotes the weight of the point x.

2 Reduction

We start by showing some preliminary results that will help us in the proof of hardness.

Consider the grid $H_{l,n}$ on the left of Fig. 1. This grid is composed by the "rows" R_i $(1 \leq i \leq l)$, alternated with the rows M_i $(1 \leq i \leq l - 1)$. The row R_i is composed by the $6n + 3$ points $\{s_i, r_{i,1}, r_{i,2}, \ldots r_{i,6n+1}, f_i\}$ (where s_i, f_i weigh w^2 and the other points weigh w), and the row M_i is composed by the $3n$ points $\{m_{i,1}, m_{i,2}, \ldots, m_{i,3n}\}$, all of weight w^2. Distances and weights are shown on the right side of Fig. 1.

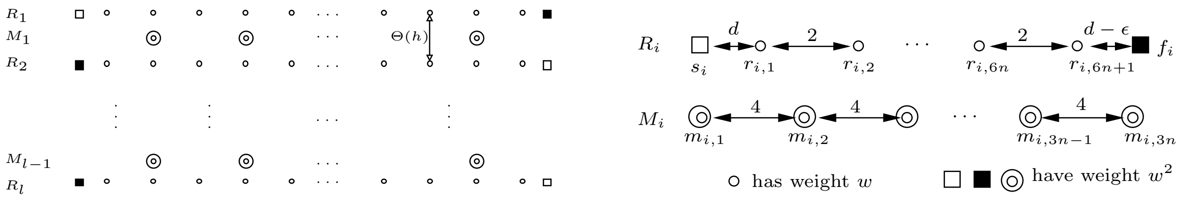


Figure 1: On the left side, the grid of points $H_{l,n}$. On the right, details of the rows

We set the following values:

$$h = w^{1/3}, \quad d = 2\sqrt{\frac{w+1}{w}}, \quad \epsilon = \frac{1}{w^2}, \quad \alpha = \frac{8}{w} - \frac{1}{w^2(w+1)}$$

Now for fixed l and n, everything is defined in terms of w. Finally we let $k = l(3n+2) + (l-1)3n$ and $L_1 = lw(6n+4)$.

Definition 4. *We define two possible $(3n+2)$-clusterings of R_i $(1 \leq i \leq l)$.*

A For $1 \leq j \leq 3n$, the j-th cluster of R_i is $\{r_{i,2j-1}, r_{i,2j}\}$. Also it has the clusters $\{s_i\}$ and $\{r_{i,6n+1}, f_i\}$.

B For $1 \leq j \leq 3n$, the j-th cluster of R_i is $\{r_{i,2j}, r_{i,2j+1}\}$. Also it has the clustes $\{s_i, r_{i,1}\}$ and $\{f_i\}$.

Definition 5. *We say that a k-clustering of $H_{l,n}$ is* nice *if each $m_{i,j}$ is a singleton cluster, and each R_i is grouped in an A-clustering or in a B-clustering.*

Lemma 6. *A nice k-clustering of $H_{l,n}$ with t rows grouped in an A-clustering costs $L_1 - t\alpha$.*

Proof. Fix a row R_i. For the clusters containing s_i and f_i, by equation (1), an A-clustering pays $\frac{w^3}{w^2+w}(d-\epsilon)^2 = 4w - \alpha$, while a B-clustering pays $\frac{w^3}{w^2+w}d^2 = 4w$. Now the lemma follows by observing that both clustering pay $(2w)(3n) = 6nw$ for the remaining clusters of the row. □

Lemma 7. *For $w = poly(n,l)$ large enough, any non-nice k-clustering of $H_{l,n}$ costs at least $L_1 + \Omega(w)$. On the other hand, any nice k-clustering of $H_{l,n}$ costs at most L_1.*

Proof. The second part of the lemma is immediate by Obs. 6. For the other part, take any non-nice clustering. If it has a cluster containing points from different rows then it would cost at least $\Omega(hw) = \Omega(w^{4/3})$. Similarly, if it has a cluster containing some $m_{i,j}$ and $m_{i,j'}$ $(j \neq j')$, then it would cost at least $\Omega(w^2)$. In both cases, for $w = poly(n,l)$ large enough, the cost is more than $L_1 + \Omega(w)$.

Now consider a non-nice clustering with each $m_{i,j}$ singleton and with no clusters containing points from different rows. We want to infer that if a row (say R_i) is not grouped as an A-clustering or a B-clustering then the cost of the clustering is at least $L_1 + \Omega(w)$.

First consider the case when R_i has a singleton cluster and the rest of the points are grouped in $3n+1$ clusters of size 2; since R_i is not nice, then the singleton cluster must be some $r_{i,j}$, while the points s_i and f_i must be in two clusters of size 2: this brings an overall cost for R_i of $2(4w) + (6n-2)w = (6n+6)w$, while a nice clustering of R_i costs at most $(6n+4)w$.

It remains to consider the case when the clustering of R_i has clusters of cardinality $m \geq 3$. Note that this allows to use more than one singleton. We claim that a cluster of cardinality m costs at least $\frac{w}{3}m(m^2-1)$. Let us conclude the proof, then we will prove the claim. Consider one of these clusters of cardinality $m \geq 3$. In a nice clustering we would have used at most $\lceil \frac{m}{2} \rceil$ clusters for the m points in the cluster, so the best we can achieve is by using the $\lceil \frac{m}{2} \rceil - 1$ "saved" clusters as singletons. Even so, for all these $m + \lceil \frac{m}{2} \rceil - 1$ points this clustering pays at least $\frac{w}{3}m(m^2-1)$. On the other hand, a nice clustering pays (a) $w(m + \lceil \frac{m}{2} \rceil - 1)$ if s_i and f_i are not among these points; (b) $w(m + \lceil \frac{m}{2} \rceil - 2) + 4w = w(m + \lceil \frac{m}{2} \rceil + 2)$ if either s_i or f_i is among these points[1]. In both cases the cost of the nice clustering is strictly better than $\frac{w}{3}m(m^2-1)$ for any $m \geq 3$.

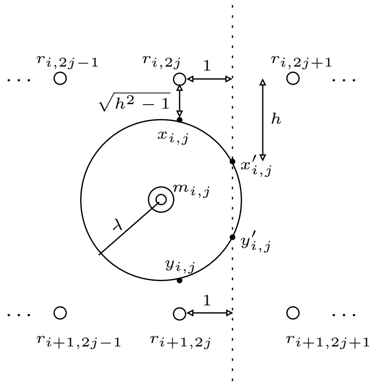
We conclude by proving our claim. Take a cluster containing $m \geq 2$ consecutive points in R_i (if the points are not consecutive the cost will be more). First we observe that we can restrict ourselves to consider the

[1] The case when both s_i and f_i are among these points would bring an even lower cost: this follows by the fact that a nice clustering has either s_i or f_i in a singleton cluster (i.e. one of them has no contribute to the cost).

case when s_i, f_i are not in the cluster. To see why, note that adding s_i (the case with f_i is analogous) to the cluster $A = \{r_{i,1}, r_{i,2}, \ldots, r_{i,j}\}$, for some $1 \le j \le m-1$, brings an incremental cost which is greater than adding $r_{i,j+1}$ to A: this follows by observing that s_i is further than $r_{i,j+1}$ from the mean of A *and* s_i weighs more than $r_{i,j+1}$. Therefore w.l.o.g. it is enough to show that the cost of $C = \{r_{i,1}, \ldots, r_{i,m}\}$ is at least $\frac{w}{3}m(m^2-1)$. Using equation (1) and observing that there are $m-i$ (unordered) pairs of points at distance $2i$ in C, we can conclude that the cost of C is

$$\frac{1}{mw}\sum_{i=0} w^2(2i)^2(m-i) = \frac{1}{mw}4w^2\left(m\frac{m(m+1)(2m+1)}{6} - \frac{m^2(m+1)^2}{4}\right) = \frac{w}{3}m(m^2-1)$$

□



Now we describe the reduction: given an instance $U = \{1, 2, \ldots, 3n\}$ and $\mathcal{C} = \{S_1, S_2, \ldots, S_l\}$ of X3C, we want to build a (decisional) instance of k-means with a set of (weighted) points $S \subseteq \mathbb{R}^2$, a certain number of clusters and a cost limit $L \in \mathbb{R}$.

We let $S = G_{l,n} \cup X$, where $G_{l,n}$ is a slight modification of $H_{l,n}$, and the set $X = \bigcup_{i=1}^{l} X_i$ depends on the collection $\mathcal{C}$.

We refer to the figure on the right to explain the details, where we set

$$\lambda = h\left(\frac{2(w^2+1)}{w(2w+1)}\right)^{1/2}$$

Note that $\lambda = \Theta(h)$. The new grid $G_{l,n}$ is identical to $H_{l,n}$ except for the fact that the points in the rows M_i are not perfectly (vertically) aligned with the points in the rows R_i. The reason is that we want the points $x_{i,j}$ and $x'_{i,j}$ (as well as the points $y_{i,j}$ and $y'_{i,j}$) to be at the same distance λ from the point $m_{i,j}$[2]. It is easy to understand that the previous results about $H_{l,n}$ apply to $G_{l,n}$ as well, since the distance between two adjacent rows in $G_{l,n}$ is again $\Theta(h)$.

Now we define the set X. The spots $x_{i,j}$, $x'_{i,j}$, $y_{i,j}$, $y'_{i,j}$ (for $1 \le i \le l-1$, $1 \le j \le 3n$) are not all points in X but only possible positions of points. Actually exactly half of these positions will be occupied: for any $1 \le i \le l-1$, $1 \le j \le 3n$, $x_{i,j} \in X_i$ iff $j \notin S_i$; $x'_{i,j} \in X_i$ iff $j \in S_i$; $y_{i,j} \in X_i$ iff $j \notin S_{i+1}$; $y'_{i,j} \in X_i$ iff $j \in S_{i+1}$. All these points have weight 1. Finally we set the number of clusters to k (recall $k = l(3n+2) + (l-1)3n$) and the cost limit to $L = L_1 + L_2 - n\alpha$, where $L_2 = 6n(l-1)h^2\frac{2w}{2w+1} = 6n(l-1)\frac{2w^{5/3}}{2w+1}$.

We now show some properties about the points in X.

Definition 8. *A cluster C is* good *for a point $z \notin C$ if adding z to C increases the cost of exactly $h^2\frac{2w}{2w+1}$.*

Lemma 9. *For any $1 \le j \le 3n$, $1 \le i \le l-1$, the following holds:*

- *The clusters $\{m_{i,j}\}$, $\{r_{i,2j-1}, r_{i,2j}\}$, and $\{r_{i,2j}, r_{i,2j+1}\}$ are good for $x_{i,j}$.*
- *The clusters $\{m_{i,j}\}$, $\{r_{i+1,2j-1}, r_{i+1,2j}\}$, and $\{r_{i+1,2j}, r_{i+1,2j+1}\}$ are good for $y_{i,j}$.*
- *The clusters $\{m_{i,j}\}$ and $\{r_{i,2j}, r_{i,2j+1}\}$ are good for $x'_{i,j}$.*
- *The clusters $\{m_{i,j}\}$ and $\{r_{i+1,2j}, r_{i+1,2j+1}\}$ are good for $y'_{i,j}$.*

Proof. By equation (1), a cluster composed of $m_{i,j}$ and a point at distance λ has a cost of $\frac{w^2\lambda^2}{w^2+1} = h^2\frac{2w}{2w+1}$.

In the other cases we start with a cluster of two points of weight w at distance 2, which has cost $2w$. W.l.o.g. assume these two points are in $(0,0)$ and $(2,0)$. Note that in all the cases we are adding a point either in $(0, \sqrt{h^2-1})$ or in $(1,h)$. In the first case, using (1), the enlarged cluster will cost $\frac{4w^2+4w+2w(h^2-1)}{2w+1} = 2w + h^2\frac{2w}{2w+1}$. In the second case, similarly, we have a cost of $\frac{4w^2+2w+2wh^2}{2w+1} = 2w + h^2\frac{2w}{2w+1}$. □

Lemma 10. *Consider any optimal k-clustering of $G_{l,n} \cup X$. Then for $w = poly(n,l)$ large enough,*

(a) the clustering induced on $G_{l,n}$ is nice;

(b) points in X are in different good clusters.

[2] Note that it must be the case that $\lambda > \frac{1}{2}d(x_{i,j}, x'_{i,j})$. This holds for w large enough since $\lambda = \Theta(h) = \Omega(w^{1/3})$ while $d(x_{i,j}, x'_{i,j})$ is upper bounded by a constant.

In addition, if there are t rows R_i grouped in an A-clustering, then this clustering costs $L_1 + L_2 - t\alpha$.

Proof. First note that we can easily find a k-clustering of $G_{l,n} \cup X$ of cost $L_1 + L_2$: start with a nice clustering of $G_{l,n}$ with all rows grouped in a B-clustering, and then for each $1 \leq i \leq l$, $1 \leq j \leq l-1$ add x_i (or x'_i) to the j-th cluster of R_i and put y_i (or y'_i) in a cluster with $m_{i,j}$. In this way, each point in X is added to a different good cluster, which leads to a total cost of $L_1 + L_2$.

Now consider an optimal k-clustering of $G_{l,n} \cup X$. To prove (a), suppose by contradiction that this clustering is not nice: then, by Lemma 7, it would cost at least $L_1 + \Omega(w)$, which for $w = poly(n, l)$ large enough is more than $L_1 + L_2$ (note that $L_2 = o(w)$).

To prove (b), note that the only way to catch up for assigning a point $x \in X$ to a non-good cluster is to increase the number of rows grouped in an A-clustering. However, by Lemma 6, even having all the rows grouped in this way would have a saving in the cost of at most $O(l\alpha)$. Since $\alpha = O(\frac{1}{w})$, while adding a point $x \in X$ to a cluster costs at least $\Omega(h^2) = \Omega(w^{2/3})$, then, for $w = poly(n, l)$ large enough, it is more convenient choosing good clusters than paying a little less by grouping the rows in A-clusterings. Finally notice that once we assign a point to a good cluster, the new (enlarged) cluster will not be good for any other point, thus any optimal clustering must assign the points in X to different good clusters. □

The following lemma proves theorem 2.

Lemma 11. *The set $G_{l,n} \cup X$ has a k-clustering of cost less or equal to L if and only if there is an exact cover $\mathcal{F} \subseteq \mathcal{C}$ for the Exact Cover by 3-sets instance.*

Proof. Fix any optimal k-clustering and suppose it costs less or equal to L. Lemma 10 allows us to define $\mathcal{F} = \{S_i : R_i \text{ is grouped in an A-clustering}\}$, and this set will contain at least n sets.

Consider any $1 \leq i \leq l$ such that R_i is grouped in an A-clustering (i.e. $S_i \in \mathcal{F}$) and consider a $j \in S_i$ which implies $x'_{i,j} \in X_i$ and $y'_{i-1,j} \in X_{i-1}$. Since R_i is grouped in an A-clustering, the point $x'_{i,j}$ cannot be in the j-th cluster of R_i because this is not a good cluster for the point. The same holds for $y'_{i-1,j}$. In other words, it holds the following claim: for any $1 \leq i \leq l, 1 \leq j \leq 3n$, if the set R_i is grouped in an A-clustering and $j \in S_i$ then the j-th cluster of R_i is not a good cluster for any points.

This means that $x'_{i,j}$ is in a cluster with $m_{i,j}$ and $y'_{i-1,j}$ is in a cluster with $m_{i-1,j}$. This implies that $y_{i,j}$ (or $y'_{i,j}$) cannot be in the cluster containing $m_{i,j}$ but has to be added to the j-th cluster of R_{i+1} which therefore has to be a good cluster since the solution is optimal. Analogously $x_{i-1,j}$ (or $x'_{i-1,j}$) cannot be in the cluster containing $m_{i-1,j}$ but has to be added to the j-th cluster of R_{i-1}. By induction the j-th cluster of all $R_{i'}$ with $i' \neq i$ has to be a good cluster for some point. The claim implies that either $R_{i'}$ is grouped as a B-clustering ($R_{i'} \notin \mathcal{F}$) or $j \notin R_{i'}$. Therefore, the sets in $\mathcal{F}$ do not overlap. Since $\mathcal{F}$ contains at least n sets, $\mathcal{F}$ is an exact cover of $\mathcal{U}$.

Conversely, suppose that the instance of Exact Cover by 3-Sets has an exact cover $\mathcal{F}$. We define now a k-clustering for $G_{l,n} \cup X$. Start putting all the points $m_{i,j}$ in different singleton clusters. For each $1 \leq i \leq l$, group R_i in an A-clustering if $S_i \in \mathcal{F}$ and in a B-clustering otherwise. For each $1 \leq j \leq 3n$, let $i(j)$ be the (unique) index such that $j \in S_{i(j)}$ and $S_{i(j)} \in \mathcal{F}$. For each $i < i(j)$ merge $x_{i,j}$ (or $x'_{i,j}$) with the j-th cluster of R_i and merge $y_{i,j}$ (or $y'_{i,j}$) with $\{m_{i,j}\}$. For each $i \geq i(j)$ merge $x_{i,j}$ (or $x'_{i,j}$) with $\{m_{i,j}\}$ and merge $y_{i,j}$ (or $y'_{i,j}$) with the j-th cluster of R_{i+1}.

Now let us compute the cost of this clustering. It is clear that we are assigning different points in $\bigcup_{i=1}^{l-1} X_i$ to different clusters. It remains to verify that all these points are assigned to good clusters. The only possibility is when, for some $1 \leq j \leq 3n$, $1 \leq i < i(j)$ (resp. $i \geq i(j)$), we merge a point $x'_{i,j}$ to the j-th cluster of R_i (resp. $y'_{i,j}$ to the j-th cluster of R_{i+1}). But notice that if $x'_{i,j} \in X_i$ (resp. $y'_{i,j} \in X_i$), then $j \in S_i$ (resp. $j \in S_{i+1}$) which implies $S_i \notin \mathcal{F}$ (resp. $S_{i+1} \notin \mathcal{F}$), which in turn implies R_i (resp. R_{i+1}) is grouped in a B-clustering, i.e. we are assigning the point to a good cluster. □

We observe that the hard instances of k-means clustering created by our reduction are such that $k = \Theta(n^\gamma)$, for some $0 < \gamma < 1$, where n is the number of points in the instance. We now prove the following result.

Theorem 12. *The k-means clustering problem is NP-hard for $k = \Theta(n^\epsilon)$, for any $\epsilon > 0$.*

Proof. Fix any $\epsilon > 0$, and take a hard instance with n points and k centers, where $k = \Theta(n^\gamma)$.

First we consider the case $\gamma < \epsilon$. In this case, we build another instance by adding n^ϵ points very far from the original instance and very far each other. Also we add n^ϵ centers. The optimal solution will use the new centers to cluster the new points, while the optimum in the original instance will not change. Therefore this is a hard instance with $m = n + n^\epsilon = \Theta(n)$ points and $k' = k + n^\epsilon = \Theta(n^\epsilon)$ centers.

Now we consider the case $\gamma > \epsilon$. Then we build another hard instance adding 1 center and $n^{\gamma/\epsilon}$ points. We put all these new points close each other but very far from the original instance, so that the new center will cluster these new points and the optimum in the original instance will not change. Therefore, we have a hard instance with $m = n + n^{\gamma/\epsilon} = \Theta(n^{\gamma/\epsilon})$ points, and $k' = k + 1 = \Theta(n^\gamma) = \Theta(m^\epsilon)$ centers. □

References

[1] M.R. Garey, D.S. Johnson, *Computers and Intractability: a Guide to the Theory of NP–completeness*, Freeman, New York, 1979.

[2] M. Inaba, N. Katoh, H. Imai. Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering. *Proceedings of 10th ACM Symposium on Computational Geometry*, pp. 332339, 1994.

[3] M. Mahajan, P. Nimbhorkar, K. Varadarajan. The Planar k-Means Problem is NP-Hard. *Lecture Notes in Computer Science* 5431: 274285, 2009.