COL870: Clustering Algorithms
Hardness of k-means clustering

Ishaan Preet Singh & Surbhi Goel

May 1, 2015

Abstract

We discuss proofs of the NP-hardness of k-means clustering, specifically for 2-means[1] and planar k-means[2][3].

1 Introduction

We can state the k-means clustering problem formally as follows -
Input: A set of points $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ and number of clusters k
Output: A partition of the points into clusters $C_1, C_2, \ldots, C_k$, and corresponding enters $\mu_1, \mu_2, \ldots, \mu_k$ minimising

$$\sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$$

In an optimal solution the centre μ_j of a cluster is simply the mean of the points in the cluster. In this case, using the fact that $\mathbb{E}\|X - Y\|^2 = 2\mathbb{E}\|X - \mathbb{E}X\|^2$, we can remove the centres from the equation. The k-means cost function now becomes -

$$\sum_{j=1}^{k} \frac{1}{2|C_j|} \sum_{x_i, x_i' \in C_j} \|x_i - x_{i'}\|^2$$

2 Hardness of 2-means Clustering

We will initially establish the hardness of k-means when $k = 2$. We will reduce 3Sat to NaeSat* and reduce that further to Generalised 2-means and then prove that our constructed problem can be embedded in the euclidean space.

2.1 Hardness of NAESAT*

NaeSat* is a special case of Not-all-equal 3Sat, and we will prove that it is hard by a reduction from 3Sat.
Input: A boolean 3CNF formula $\phi(x_1, \ldots, x_n)$ such that

1. Each clause contains exactly 3 literals.
2. Each pair appears together in a clause at most twice and if appears twice then once as $\{x_i, x_j\}$ or $\{\overline{x}_i, \overline{x}_j\}$ and once as $\{x_i, \overline{x}_j\}$ or $\{\overline{x}_i, x_j\}$

Output: *true* if $\exists$ an assignment in which each class has one or two satisfied literals (i.e. not all *equal*) else $false$.

Reduction from 3SAT

We are given an input $\phi(x_1, \ldots, x_n)$ to 3SAT.

1. Construct intermediate ϕ': For a variable x_i that occurs in k clauses, create k variables $x_{i1}, x_{i2}, \ldots, x_{ik}$ and replace each occurrence of x_i by one of the new variables. Also, add clauses $(\overline{x_{i1}} \vee x_{i2}), (\overline{x_{i2}} \vee x_{i3}), \ldots (\overline{x_{ik}} \vee x_{i1})$ which ensure that all the new variables for x_i have the same value. ϕ' is obviously equivalent to ϕ and 2 variables never occur together in a clause more than once.

2. Constructing ϕ'': In ϕ', let the number of 2 variable clauses be m and the number of 3 variable clauses be m'. Create new variables $s_1, s_2, \ldots, s_m$ and $f_1, f_2, \ldots, f_{m+m'}$ and f.
Given the jth 3 literal clause, $(\alpha \vee \beta \vee \gamma)$ replace it with $(\alpha \vee \beta \vee s_j)$ and $(\overline{s}_j \vee \gamma \vee f_j)$.
Given the jth 2 literal clause $(\alpha \vee \beta)$ replace it with $(\alpha \vee \beta \vee f_{m+j})$.
Also, add clauses $(\overline{f}_1 \vee f_2 \vee f_3), (\overline{f}_2 \vee f_3 \vee f_4), \ldots, (\overline{f}_{m+m'} \vee f_1 \vee f)$ which ensure that all f_is have the same value (if f is false).
Now, all clauses have 3 literals each, all f_is must have the same value in a satisfying assignment (if f is false), and only (f_i, f) occur more than once in a pair and they satisfy the required conditions for NaeSat*.

Lemma 1 *ϕ' is satisfiable if and only if ϕ is not-all-equal satisfiable.*

Proof. If ϕ' is satisfiable, keep the same values of variables for ϕ', set all f_is and f as $false$, and for a 3 variable clause $(\alpha \vee \beta \vee s_j)$ set s_j as $false$ if both α and β are $false$, then set s_j to true, satisfying the first clause, while $(\overline{s}_j \vee \gamma \vee f_j)$ is satisfied because γ must be true. Else, set s_j to false. The first clause is already satisfied and is not-all-equal because of s_j. The second is satisfied because of $\overline{s}_j$ and is not-all-equal because of f_j. The case of 2 literal clauses is simple because all f_is in them are false, while at least one of α or β must be true.
Now, suppose ϕ'' is not-all-equal satisfiable. Note that if an assignment of variable is not-all-equal satisfiable, we can flip all assignments and the satisfiability remains true. This is because at least one of the variables in every clause was false (and not all were false), meaning not all will be true, and at least one will be. Let us assume that f is false (if it isn't flip all assignments). Now, all f_is are equal. If they aren't false, flip all assignments. This means all f_is are now false. Hence, all 2 literal clauses are now satisfied in ϕ'. In the 3 literal clauses, since f_j is false, at least one of α, β or γ must be true meaning $(\alpha \vee \beta \vee \gamma)$ is satisfied. ∎

2.2 Hardness of Generalised 2-means

In the generalised k-means problem instead of using the Euclidean distances between points, we assume that we are given an nxn distance matrix D and we try and cost function -

$$\sum_{j=1}^{2} \frac{1}{2|C_j|} \sum_{x_i, x_i' \in C_j} D_{ii'}$$

Reduction from NAESAT*

We are given an instance of NaeSat* with input $\phi(x_1, \ldots, x_n)$ and we construct a generalised 2-means problem with $2n$ points, with points corresponding to $x_1, x_2, \ldots, x_n, \overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n$. We define $\alpha\ \beta$ as implying that α and β occur together in a clause or $\overline{\alpha}$ and $\overline{\beta}$ occur together. Two different clauses can not imply that $\alpha\ \beta$ because of our input restrictions on pairs. Define -

$$D_{\alpha\beta} = \begin{cases} 0 & \text{if } \alpha = \beta \\ 1 + \Delta & \text{if } \alpha = \overline{\beta} \\ 1 + \delta & \text{if } \alpha \sim \beta \\ 1 & \text{otherwise} \end{cases}$$

Here, $0 < \delta < \Delta < 1$ and $4\delta m < \Delta \leq 1 - 2\delta n$ are constraints on δ and Δ, where m is the number of clauses. $\delta = 1/(5m + 2n)$ and $\Delta = 5\delta m$ is one such valid setting.

Lemma 2 *If ϕ is a satisfiable instance of* NaeSat**, then the above construction admits a generalised 2-means clustering of cost $c(\phi) = n - 1 + 2\delta m$.*

Proof. Take all variables assigned true in one cluster and all variables with value false in the other. Each cluster must have n points. Within a cluster there are no variables such that $\alpha = \overline{\beta}$. Hence, the distances are either 1 or $1 + \delta$. Each clause is necessarily split among the clusters, because if it had all 3 of its variables in one cluster they would either all be true (not NaeSat*) or not be satisfiable. Hence, each clause has at least one variable in C_1 and one in C_2. This means that it contributes either one pair of α β points to C_1 or one pair to C_2. Hence, each clause results exactly one pair of such points, meaning there are m such points.

$$c(\phi) = \frac{1}{2n} \sum_{i,i' \in C_1} D_{ii'} + \frac{1}{2n} \sum_{i,i' \in C_2} D_{ii'} = 2.\frac{1}{n}\left(\binom{n}{2} + m\delta\right) = n - 1 + 2\delta$$

This is true because for every pair of points the distance will either be 1 or $1 + \delta$ and it will be $1 + \delta$ m times. ■

Lemma 3 *For any 2-clustering C_1, C_2, if C_1 contains both a variable and it's negation, then the cost is at least $c(\phi)$.*

Proof. Let C_1 have n' points. Since all distances are at least 1 and C_1 contains a pair of points with distance $1 + \Delta$, the cost of the clustering is at least

$$\frac{1}{n'}\left(\binom{n'}{2} + \Delta\right) + \frac{1}{2n - n'}\binom{2n - n'}{2} = n - 1 + \frac{\Delta}{n'} \geq n - 1 + \frac{\Delta}{2n} \geq c(\phi)$$

■

Lemma 4 *If D admits a 2-clustering of cost $\leq c(\phi)$, then ϕ is a satisfiable instance of* NaeSat*.

Proof. By the previous lemma, neither of the clusters have both a variable and a negation, implying that they are split equally across the clusters. Hence, $|C_1| = |C_2| = n$. Now, cost of the clustering can be written as -

$$\frac{2}{n}\left(\binom{n}{2} + \delta \sum_{clauses} (1 \text{ if clause is split between } C_1 \text{ and } C_2; 3 \text{ otherwise})\right).$$

For the cost to be $\leq c(\phi)$, all of the clauses should be split between C_1 and C_2. If a clause had all 3 variable in one cluster then it would form 3 pairs which would make the cost more than $c(\phi)$, as for $c(\phi)$ each clause only contributed one such pair. Hence, setting all variables in C_1 as true and the rest as false will mean ϕ is NaeSat*. ■

2.3 Embeddability of the Construction

We will now show that the D matrix we constructed is 'embeddable' meaning that there exists corresponding points $x_\alpha \in \mathbb{R}^{2n}$ such that $D_{\alpha\beta} = \|x_\alpha - x_\beta\|$ for all α, β. To prove this we will use the following theorem from [4]

Theorem 5 *Let H denote the matrix $I - (1/N)\mathbf{1}\mathbf{1}^T$. An NxN matrix is embeddable if and only if $-HDH$ is positive semi definite.*

Lemma 6 *An NxN matrix is embeddable if and only if $u^T Du \leq 0$ for all $u \in \mathbb{R}^n$ such that $u.\mathbf{1} = 0$.*

Proof. The range of $v \to Hv$ is $\{u \in \mathbb{R}^n : u.\mathbf{1} = 0\}$. Hence,

$$\begin{aligned} -HDH \text{ is positive semidefinite} &\iff v^T HDHv \leq 0 \text{ for all } v \in \mathbb{R}^n \\ &\iff u^T Du \leq 0 \text{ for all } u \in \mathbb{R}^n \text{ such that } u.\mathbf{1} = 0 \end{aligned}$$

■

Lemma 7 *$D(\phi)$ is embeddable*

Proof. $D(\phi)$ is a $2n$x$2n$ matrix constructed from $\phi(x_1, x_2, \ldots, x_n)$, with the first n indices corresponding to $x_1, x_2, \ldots, x_n$ and the next n corresponding to $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n$.
Pick any $u \in \mathbb{R}^2 n$ such that $u.\mathbf{1} = 0$. Define u^+ as the first n coordinates of u and u^- as the last n coordinates.

$$\begin{aligned}
u^T D u &= \sum_{\alpha,\beta} D_{\alpha\beta} u_\alpha u_\beta \\
&= \sum_{\alpha,\beta} u_\alpha u_\beta - \sum_{\alpha} u_\alpha^2 + \Delta \sum_{\alpha} u_\alpha \overline{u}_\alpha + \delta \sum_{\alpha,\beta} u_\alpha u_\beta (1 \text{ if } \alpha \sim \beta, 0 \text{ otherwise}) \\
&\leq \left(\sum_{\alpha} u_\alpha \right)^2 - \|u\|^2 + 2\Delta(u^+ u^-) + \delta \sum_{\alpha,\beta} |u_\alpha||u_\beta|
\end{aligned}$$

Now, $\sum_{\alpha} u_\alpha = 0$ since $u.\mathbf{1} = 0$ and we can use $(a+b)^2 > 0$ for the third and fourth term. Hence,

$$\begin{aligned}
u^T D u &\leq -\|u\|^2 + \Delta(\|u^+\|^2 + \|u^-\|^2) + \delta \left(\sum_{\alpha} |u_\alpha| \right)^2 \\
&\leq -(1-\Delta)\|u\|^2 + 2\delta \|u\|^2 n
\end{aligned}$$

The last step used the Cauchy-Shwartz inequality on the last term. Now, this quantity is always negative since $2\delta n \leq 1 - \Delta$. ■

3 Hardness of Planar k-means Clustering

In this section, we restate the proof of the hardness of k-means clustering for $d = 2$ dimensions given in ([3] uses a different reduction to prove the same). The hardness result holds for $k = \Theta(n^\epsilon)$, for any $\epsilon > 0$, where n is the number of points and k is the number of clusters. We use the decisional version of weighted k-means clustering problem. W.l.o.g. we can replace a point x of weight w with w distinct points within very close distance of x.

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

It is clear that the above problem is in NP as any solution can be verified in randomized polynomial time. We will prove that this problem is in fact NP-complete for $d = 2$ by reduction from Exact Cover by 3-Sets (X3C) which is known to be NP-complete.

Definition 2 *Given a finite set U containing exactly $3n$ elements and a collection $\mathcal{C} = \{S_1, S_2, ..., S_l\}$ of subsets of U each of which contains exactly 3 elements, are there n sets in $\mathcal{C}$ such that their union is U?*

3.1 Preliminary Results

Consider the grid $H_{l,n}$ as shown in the figure. The grid consists of l rows indexed by R_i $(1 \leq i \leq l)$ alternated with $l-1$ rows indexed by M_i $(1 \leq i \leq l-1)$. Each R_i consists of $6n+3$ points whereas row M_i consists of $3n$ points. The positions, labels and weights are as indicated in the figure.

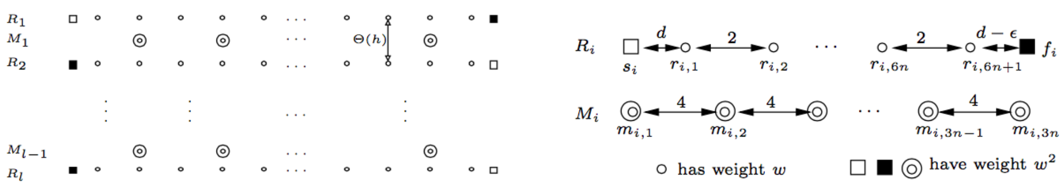


Figure 1: On the left side, the grid of points $H_{l,n}$. On the right, details of the rows

Set the following values:

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

Definition 3 *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 clusters* $\{s_i, r_{i,1}\}$ *and* $\{f_i\}$*.*

Definition 4 *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 8 *A nice* k*-clustering of* $H_{l,n}$ *with* t *rows grouped in an* A*-clustering costs* $L_1 - t\alpha$ *where* $L_1 = lw(6n+4)$*.*

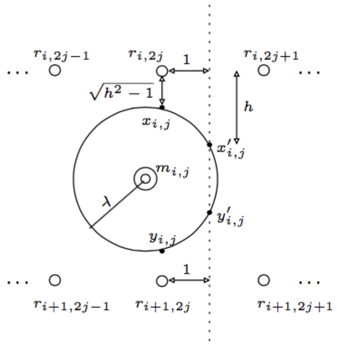
Proof. Clustering A and B differ in terms of cost due to the clusters $\{r_{i,6n+1}, f_i\}$ and $\{s_i, r_{i,1}\}$ respectively since the singletons do not add to the cost and the remaining $3n$ clusters consist of 2 points of weight w each separated by distance 2. Cost of the latter by simple calculation works out to be $(2w)(3n) = 6nw$. Due to the former different clusters, A pays $\frac{w^3}{w^2+w}(d-\epsilon)^2 = 4w - \alpha$ and B pays $\frac{w^3}{w^2+w}d^2 = 4w$. Hence, the total cost is $t(4w-\alpha) + (l-t)(4w) + 6nwl = L_1 - t\alpha$. ■

Lemma 9 *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 statement follows from the above lemma as the cot of a nice clustering is bounded by L_1. For the first part, we will consider the following cases:
Case 1: Cluster contains points from different rows.
Since the rows are separated by distance $\Theta(h)$, the cost ill be at least $\Omega(hw) = \Omega(w^{4/3})$.
Case 2: Cluster contains 2 points from row M_i**.**
The cost of such a cluster will be least when the two points are consecutive and even for this case the cost works out to be $8w^2$.
In both cases, for $w = poly(n, l)$ large enough, the cost is more than $L_1 + \Omega(w)$ as L_1 is linear in w. This implies that each $m_{i,j}$ is a singleton and no element from different rows are in the same cluster.
Case 3: R_i **contains a singleton cluster and rest grouped in** $3n+1$ **pairs.**
Since R_i is not nice, the singleton must be some $r_{i,j}$ while the points s_i and f_i are in 2 size clusters. The overall cost for this arrangement simply works out to be $4w + 4w - \alpha + (3n-1)(2w) = (6n+6)w - \alpha$ while a nice clustering costs at most $(6n+4)w$. For large w, α is very small, cost of this non-nice clustering exceeds that of a nice clustering by a factor of w.
Case 4: R_i **is not nice and contains clusters of cardinality** $m \geq 3$**.**
Consider a cluster of cardinality $m \geq 3$. In a nice clustering, these m points would have used at most $\lceil \frac{m}{2} \rceil$ clusters (assume that the m points are continuous for least cost), so the best we can do is by using $\lceil \frac{m}{2} \rceil - 1$ singletons. Using simple calculations, we can show that a cluster of cardinality m costs at least $\frac{w}{3}m(m^2-1)$. So this case would cost at least $\frac{w}{3}m(m^2-1)$. Whereas a nice clustering would cost at most $w(m + \lceil \frac{m}{2} \rceil - 1)$ if s_i or f_i are not among the points else $w(m + \lceil \frac{m}{2} \rceil - 1) + 4w = w(m + \lceil \frac{m}{2} \rceil + 2)$ (consider B clustering as it has higher cost and we are upper bounding). In both these cases, the cost of non nice is strictly worse than nice clustering. ■

3.2 Reduction

In this section we will describe the main reduction i.e. we will build a decisional instance of weighted k-means with a certain k and a cost limit $L \in \mathbb{R}$ for a given instance of X3C.



To do so we use $G_{l,n}$, a slight modification of above mentioned $H_{l,n}$ as shown in figure. The main difference is that the position of each $m_{i,j}$ is not perfectly vertically aligned as before. Trivially, this modification preserves all the lemmas from the previous section (distance between the two rows remains same). In the figure,

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

We define set $S = G_{l,n} \cup X$ where $X = \bigcup_{i=1}^{l-1} X_i$ and it depends on the collection $\mathcal{C}$ of the X3C problem. The points in the figure $x_{i,j}, x'_{i,j}, y_{i,j}, y'_{i,j}$ for each i, j are possible points in X. Their presence in X is governed by the following rules:

- $x_{i,j} \in X_i$ 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 $j \in S_{i+1}$

We will solve the k-means problem on the defined S with k as in previous section. The intuition for the reduction is that the arrangement of the clusters defines the sets to choose in the X3C problem and the added points take care of the disjoint property of the selected sets. To formally show the reduction, we will define some properties about the points in X.

Definition 5 *A cluster C is good for a point $z \notin C$ if adding z to C increases the cost by exactly $h^2\frac{2w}{2w+1}$.*

Lemma 10 *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 $\{r_{i+1,2j}, r_{+1i,2j+1}\}$ 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. The result is straightforward through simple calculations. ■

Lemma 11 *Consider any optimal k-clustering of $G_{l,n} \cup X$. Then for $w = poly(n,l)$ large enough,*

1. *the clustering induced on $G_{l,n}$ is nice;*
2. *points in X are in different good clusters.*

In addition, if there are t rows R_i grouped in an A-clustering, then this clustering costs $L_1 + L_2 - t\alpha$ where $L_2 = 6n(l-1)h^2\frac{2w}{2w+1}$.

Proof. Using lemma 1 and 3, we can define a clustering for S with cost $L_1 + L_2$. To do so, we start with a nice clustering of $G_{l,n}$ with all rows grouped in B-clustering (cost is L_1) and for each $x_{i,j}(x'_{i,j})$, we add it to the j-th cluster of R_i and put $y_{i,j}(y'_{i,j})$ to the cluster $\{m_{i,j}\}$, both are good clusters for the corresponding points. Since all points are added to good clusters, the cost increase from these points is exactly L_2 resulting in the total cost of $L_1 + L_2$. Thus, the optimal clustering must have cot $\le L_1 + L_2$. By lemma 1, cost of any non-nice clustering of $G_{l,n}$ is at least $L_1 + \Omega(w)$, which for large w exceeds $L_1 + L_2$. This proves 1.

Now, we need to show that the points are in different good clusters. If we assign a point to a non-good cluster, we will have to compensate by increasing the number of rows in A-clustering. By lemma 1, we can decrease the cost by maximum $l\alpha$ (note that α is $O(\frac{1}{w})$). Adding a point x to a cluster costs at least $\Omega(h^2) = \Omega(w^{2/3})$ (from figure), for large w, cost of assigning to a non-good cluster can not be compensated resulting in a higher cost clustering. Thus, each $x \in X$ is assigned to a good cluster. Also, a cluster does not remain good after adding a point, implying that points in X are assigned to different clusters. Cost of this clustering is direct from lemma 1 and 3. ■

Lemma 12 *The set $S = G_{l,n} \cup X$ has a k-clustering of cost less or equal to $L = L_1 + L_2 - n\alpha$ if and only if there is an exact cover $\mathcal{F} \subseteq \mathcal{C}$ for the Exact Cover by 3-sets instance.*

Proof. We give an overview of the proof without giving complete details. Refer [2] for details.

Consider an optimal k-clustering of S with cost less or equal to L. The optimal clustering must be of the form as in lemma 4. This lets us define $\mathcal{F} = \{S_i : R_i \text{ is grouped in an } A\text{-clustering}\}$ such that $|\mathcal{F}| \geq n$. To show this to be an exact cover, we claim that for i such that $S_i \in \mathcal{F}$ and $j \in S_i$, for all $i' \neq i$, $R_{i'}$ is grouped as a B-clustering or $j \notin S_{i'}$. Assuming this to be true, all sets in $\mathcal{F}$ are disjoint, thus union of n of these is S and $\mathcal{F}$ is an exact cover. To prove the claim, the high level idea is to use induction to show that the jth-cluster of each $R_{i'}$ is a good cluster which implies the claim. To do this we consider the possible clusters the points $x_{i',j}(x_{i',j})$ and $y_{i',j}(y'_{i',j})$ have to be assigned to given $x'_{i,j}$ is assigned to $\{m_{i,j}\}$ and $y'i-1, j$ is assigned to $\{m_{i-1,j}\}$.

For the converse, we construct the clustering by assigning R_i as an A-clustering if $S_i \in \mathcal{F}$ and B-clustering otherwise. We then assign points appropriately to good clusters (depending on the index of the sets in which each element belongs to). ■

In the above analysis, we have $k = \Theta(n^\gamma)$ for some $0 < \gamma < 1$. The last thing that remains to be done is to generalize this to any $\epsilon > 0$.

Theorem 13 *The k-means clustering problem is NP-hard for $k = \Theta(n^\epsilon)$, for any $\epsilon > 0$.*

Proof. Fix an $\epsilon > 0$ and take a hard instance with n points and k centers where $k = \Theta(n^\gamma)$.
Case 1 ($\gamma < \epsilon$): Construct a new instance with n^ϵ points added very far from the original problem as well as from each other. Adding n^ϵ centers gives the optimal solution as the optimal for the original plus each of the added points. Thus, this is a hard problem for $m = n + n^\epsilon = \Theta(n)$ points and $k' = k + n^\epsilon = \Theta(n^\epsilon)$ centers.
Case 2 ($\gamma > \epsilon$): Construct a new instance with $n^{\gamma/\epsilon}$ points added very far from the original problem and very close to each other. Adding 1 center gives the optimal solution as the optimal for the original plus one with the cluster of new points. Thus, this is a hard problem for $m = n + n^{\gamma/\epsilon} = \Theta(n^{\gamma/\epsilon})$ points and $k' = k + 1 = \Theta(n^\gamma) = \Theta(m^\epsilon)$ centers.

■

References

[1] S. Dasgupta. The hardness of k-means clustering. Technical Report CS2007-0890, University of California, San Diego, 2007.

[2] Andrea Vattani. The hardness of k-means clustering in the plane. manuscript, 2009.

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

[4] I.J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44:522?553, 1938.