The Planar k-means Problem is NP-hard[☆]

Meena Mahajan[a], Prajakta Nimbhorkar[a], Kasturi Varadarajan[b]

[a] *The Institute of Mathematical Sciences, Chennai 600 113, India.*
[b] *The University of Iowa, Iowa City, IA 52242-1419 USA.*

Abstract

In the k-means problem, we are given a finite set S of points in $\Re^m$, and integer $k \geq 1$, and we want to find k points (centers) so as to minimize the sum of the square of the Euclidean distance of each point in S to its nearest center. We show that this well-known problem is NP-hard even for instances in the plane, answering an open question posed by Dasgupta [7].

1. Introduction

In the k-means problem, we are given a finite set S of points in $\Re^m$, and integer $k \geq 1$, and we want to find k points (centers) so as to minimize the sum of the square of the Euclidean distance of each point in S to its nearest center. This is a well-known and popular clustering problem that has also received a lot of attention in the algorithms community.

Lloyd [17] proposed a very simple and elegant local search algorithm that computes a certain local (and not necessarily global) optimum for this problem. Har-Peled and Sadri [11] and Arthur and Vassilvitskii [5, 4] examine the question of how quickly this algorithm and its variants converge to a local optimum. Lloyd's algorithm also does not provide any significant guarantee about how well the solution that it computes approximates the optimal solution. Ostrovsky et al. [20] and Arthur and Vassilvitskii [6] show that randomized variants of Lloyd's algorithm can provide reasonable approximation guarantees.

The k-means problem has also been studied directly from the point of view of approximation algorithms. There are polynomial time algorithms that compute a constant factor approximation to the optimal solution; see for instance the local search algorithm analyzed by Kanungo et al. [13]. If k, the number of centers, is a fixed constant, then the problem admits polynomial-time approximation schemes [8, 14]. If both k and the dimension m of the input are fixed, the problem can be solved exactly in polynomial time [12].

[☆] Part of the work by the third author was done when visiting The Institute of Mathematical Sciences, Chennai. He was also supported by NSF CAREER award CCR 0237431.

Email addresses: meena@imsc.res.in (Meena Mahajan), prajakta@imsc.res.in (Prajakta Nimbhorkar), kvaradar@cs.uiowa.edu (Kasturi Varadarajan)

Drineas et al. [9], Aloise et al. [2], and Dasgupta [7] show that the k-means problem is NP-hard when the dimension m is part of the input even for $k = 2$. However, to the best of our knowledge, there is no known NP-hardness result when the dimension m is fixed and k, the number of clusters, is part of the input. Dasgupta [7] raises the question of whether k-means is hard in the plane.

In this paper, we establish the NP-hardness of the k-means problem in the plane. Thus the main result of the paper is the following:

Theorem 1. *Given a finite set $S = \{p_i, p_2, \ldots, p_n\}$ of points with rational coordinates in $\Re^2$, an integer $k \geq 1$, and a bound $R \in \Re$ which is a rational number, it is NP-hard to determine if there exist k centers $\{c_1, \ldots, c_k\}$ anywhere in $\Re^2$ such that*

$$\sum_{i=1}^{n} \left(\min_{1 \leq j \leq k} [d(p_i, c_j)]^2 \right) \leq R$$

Our proof uses a reduction from the planar 3SAT problem [16] and is inspired by a construction used by Megiddo and Supowit [19] in the context of showing the NP-hardness of the k-center and the k-median problem. Given an instance $\mathcal{I}$ of planar 3-SAT, we construct a planar graph from $\mathcal{I}$, embed it in an integer grid using the technique of [1], and construct the k-means instance from this grid embedding. (See Figure 4.)

While clustering problems generally tend to be NP-hard even in the plane, there are surprising exceptions – the problem of covering a set of points by k balls so as to minimize the sum of the radii of the balls admits a polynomial time algorithm if we use L_1 balls, and a $(1 + \varepsilon)$-approximation algorithm that runs in time polynomial in the input size and $\log \frac{1}{\varepsilon}$ for the usual Euclidean balls [10].

The rest of this article is organized as follows. In Section 2, we define the problem formally and state some useful properties of the optimal clustering. In Section 3, we describe our reduction from planar 3SAT to the k-means in the plane.

2. Preliminaries

The problem of k-means clustering is defined as follows:

Definition 2. *Given a set of n points $S = \{p_1, \ldots, p_n\}$ in $\mathbb{R}^m$, find a set of k points $B = \{b_1, b_2, \ldots, b_k\} \subset \mathbb{R}^m$ such that*

$$\sum_{i=1}^{n} [d(p_i, B)]^2$$

is minimized. This minimum value is denoted $Opt(S, k)$.

Here $d(p_i, B)$ is the Euclidean distance from p_i to the nearest point in B; $d(p_i, B) = \min_{1 \leq j \leq k} d(p_i, b_j)$.

We consider the problem for $m = 2$, and refer to it as *planar k-means*.

Choosing the set B of k centers fixes a clustering $\mathcal{C}$ of the points in S, with each point going to its nearest center (breaking ties arbitrarily). On the other hand, if a set $C \subseteq S$ is known to form a cluster, then the center of the cluster is uniquely determined as the centroid of the points in C. Thus we can talk of the cost of a cluster $C = \{p_1, \ldots, p_m\}$ and a clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$:

$$\mathrm{Opt}(C, 1) = \mathrm{Cost}(C) = \min_b \sum_{i=1}^{m} [d(p_i, b)]^2$$

$$\mathrm{Cost}(\mathcal{C}) = \sum_{j=1}^{k} \mathrm{Cost}(C_j)$$

Thus $\mathrm{Opt}(S, k)$ is the minimum, over all clusterings $\mathcal{C}$ of S into k clusters, of $\mathrm{Cost}(\mathcal{C})$.

We show the NP-hardness of planar k-means by a reduction from planar 3-SAT [16].

Definition 3. *([16]) Let F be a 3-CNF formula with variables $\{v_1, \ldots, v_n\}$ and clauses $\{c_1, \ldots, c_m\}$. We call $G(F)$=(V, E) the graph of F, where*

$$\begin{aligned} V &= \{v_i | 1 \le i \le n\} \cup \{c_j | 1 \le j \le m\} \\ E &= E_1 \cup E_2 \text{ where} \\ E_1 &= \{(v_i, c_j) | v_i \in c_j \text{ or } \bar{v_i} \in c_j\} \\ E_2 &= \{(v_j, v_{j+1}) | 1 \le j < n\} \cup \{(v_n, v_1)\} \end{aligned}$$

If $G(F)$ is a planar graph, F is called a planar 3-CNF formula. The planar 3-SAT problem is to determine whether a given planar 3-CNF formula F is satisfiable.

We note that our reduction, in fact, requires only the graph (V, E_1) to be planar. (Some of the literature in fact refers to this sub-graph as the graph of F, but we follow the convention from [16].)

Henceforth throughout this note, we use the term distance to mean square of Euclidean distance. That is, $\mathrm{dist}(p, q) = [d(p, q)]^2$. We will be explicit when deviating from this convention.

We use the following well known or easily verifiable facts about the k-means problem [12, 9].

Proposition 4. 1. *The cost of a cluster of points is half the average sum of distances from a point to the other points in the cluster:*

$$\mathit{Cost}(S) = \frac{1}{2|S|} \sum_{p \in S} \sum_{q \in S; q \neq p} \mathit{dist}(p, q)$$

In other words, if $S = \{p_1, p_2, \ldots, p_n\}$, then

$$\mathit{Cost}(S) = \frac{1}{|S|} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathit{dist}(p_i, p_j)$$

In the following, we will use this form as the definition for the cost of a cluster.

2. *If, in an instance of the k-means problem, the given points form a multiset, then we say that a clustering is multiset-respecting if it puts all points at the same location into the same cluster.*
   Every instance of the k-means problem has a multiset-respecting optimal clustering.
3. *Let S be a multiset instance of the k-means problem, and let S' be the instance obtained by adding a point p to S. Then $\forall k$, $Opt(S,k) \leq Opt(S',k)$.*
4. *In particular, adding a point to a cluster cannot decrease the cost of that cluster; $Cost(S) \leq Cost(S \cup \{p\})$.*
5. *If clustering $\mathcal{C}'$ refines clustering $\mathcal{C}$ (that is, every cluster in $\mathcal{C}$ is the union of some clusters in $\mathcal{C}'$), then $Cost(\mathcal{C}') \leq Cost(\mathcal{C})$.*

3. Reduction from planar 3-SAT to k-means

Let F be the given planar 3SAT instance with n variables and m clauses. We construct an instance I of planar k-means corresponding to F. We list the required properties of this instance in Section 3.1. In Section 3.3, we describe a layout which indeed satisfies these properties, and also prove the correctness of the reduction. The reduction may introduce irrational coordinates in the resulting k-means instance. Rounding these irrational coordinates to sufficiently close rational points is described in Section 3.4.

3.1. Properties of the layout

The corresponding k-means instance I we construct will satisfy the following:

1. Corresponding to each variable x_i, there is a simple circuit s_i in the plane, with an even number of vertices marked on it. At each vertex on such a circuit, M copies of a point are placed. The circuits for different variables do not intersect.
   For each circuit, its vertices can be partitioned into pairs of adjacent vertices in two ways. We associate one of them (chosen arbitrarily) with the assignment $x_i = 1$ and the other with $x_i = 0$. We call the first pairing the 'true matching' and the other pairing the 'false matching'.
2. Let u, v be any two distinct vertices taken from any of the circuits (not necessarily the same circuit). If u and v are adjacent on some circuit, then the distance between them is β. Otherwise, the distance between them is at least 2β.
3. There is a point p_j corresponding to every clause C_j. If $x_i \in C_j$ ($\bar{x_i} \in C_j$) then there is a unique nearest edge (u, v) on the true (respectively false) matching of the circuit s_i such that p_j is equi-distant from u and v. It is at distance α from the midpoint of uv, and hence at a distance $\alpha + \frac{\beta}{4}$ from u and v. All vertices other than the endpoints of these nearest edges (two per literal in the clause, so at most six) are at a distance at least $\alpha + \frac{5\beta}{4}$ from p_j.

Clause points p_j and p_l, for $l \neq j$, are at distance at least θ from each other.

4. The instance I consists of all the clause points, and M copies of a point at each vertex on each circuit s_i. The parameters satisfy

$$M \geq \frac{6\alpha m}{\beta} \qquad\qquad \theta \geq 2(M+1)\alpha m$$

5. The value of k is given by

$$k = \sum_{i=1}^{n} \frac{|s_i|}{2}$$

We ensure that the optimal k-means clustering puts the points in each circuit s_i into $\frac{|s_i|}{2}$ clusters by dividing them into either true pairs or false pairs. (Thus these clusters contain $2M$ points.) Every clause point p_j has at most three pairs of point locations at distance α from itself. It is clustered with one of these pairs if that pair forms a cluster in the circuit s_i. Otherwise, the optimal clustering puts p_j along with some pair of point locations that forms a cluster in the circuit it appears in. In particular, if x_i is assigned a value 1, then in the corresponding k-means clustering, points of s_i are clustered according to the true matching, otherwise they are clustered according to the false matching. Similarly, if the assignment to a variable x_i satisfies a clause c_j, then the clause point p_j is at distance $\alpha + \frac{\beta}{4}$ from the vertices of a cluster in s_i, otherwise it is at distance strictly greater than $\alpha + \frac{\beta}{4}$ from at least one vertex in every cluster in s_i.

We need to show that

1. A layout satisfying the above properties gives a correct reduction from planar 3-SAT to planar k-means. This is done in Section 3.2 below.
2. The layout is indeed possible for some choice of α, β, θ, M, and can be obtained in polynomial time. This is done in two stages: a layout with irrational coordinates is described in Section 3.3, and in Section 3.4 we describe how to eliminate the irrational coordinates.

3.2. Correctness of the Reduction

Consider clustering of only circuit points into k non-empty clusters.

Lemma 1. 1. *Clustering the circuit points into consecutive pairs (i.e. into the true or the false matching for each variable) has cost* $\frac{kM\beta}{2}$.

2. *Any other multiset-respecting clustering of circuit points has cost at least* $\frac{kM\beta}{2} + \frac{M\beta}{3}$.

Proof: Let A be any matching-based k-means clustering of the circuit points. Then using Proposition 4(1) we can see that $\text{Cost}(A) = \frac{kM\beta}{2}$.

Let B be some multiset-respecting clustering that does not correspond to a matching on the circuits. By the size of a cluster, we mean the number of distinct vertices (and hence all M points at that vertex) in it.

If the largest cluster in B has 2 vertices, then every cluster is a pair, and at least one pair is not consecutive on any circuit. Hence

$$\text{Cost}(B) \geq \frac{(k-1)M\beta}{2} + \frac{M^2(2\beta)}{2M} = \frac{kM\beta}{2} + \frac{M\beta}{2}$$

satisfying the claimed bound.

So now assume that B has some larger clusters too. Let B contain p clusters of sizes $l_1, \ldots, l_p$ more than 3 each, q clusters of size 3 each, r clusters of size 2 each, and s clusters of size 1 each. Then we have the following:

$$p + q + r + s = k \tag{1}$$

$$\sum_{i=1}^{p} l_i + 3q + 2r + s = 2k \tag{2}$$

Subtracting twice the first equation from the second, and using $p = \sum_{i=1}^{p} 1$, we get

$$s = \sum_{i=1}^{p} (l_i - 2) + q \tag{3}$$

For a cluster C of size $l \geq 4$, the best possible situation is that l pairs within the cluster are edges on some circuit. Thus the cost of such a cluster is at least

$$\text{Cost}(C) \geq \frac{1}{lM}\left[lM^2\beta + \left(\binom{l}{2} - l\right) M^2 2\beta\right] = (l-2)M\beta$$

Similarly, in a cluster of size 3, at most two pairs can be edges on a circuit (the circuits are of even length), so the cost is at least $4M\beta/3$.

The cost of the (p, q, r, s) clustering B thus satisfies:

$$\begin{aligned}
\text{Cost}(B) &\geq M\beta\left[\sum_{i=1}^{p}(l_i - 2) + \frac{4q}{3} + \frac{r}{2}\right] \\
&= \frac{M\beta}{2}\left[s + r + q + \frac{2q}{3} + \sum_{i=1}^{p}(l_i - 2)\right] \qquad \text{from } (\textit{Equation } 3) \\
&\geq \frac{M\beta}{2}\left[s + r + q + p + \frac{2q}{3} + p\right] \qquad \because l_i - 2 \geq 2 \\
&= \frac{M\beta}{2}\left[k + p + \frac{2q}{3}\right] \qquad \text{from } (\textit{Equation } 1) \\
&\geq \frac{kM\beta}{2} + \frac{(p+q)M\beta}{3} \\
&\geq \frac{kM\beta}{2} + \frac{M\beta}{3} \qquad \because p + q \geq 1
\end{aligned}$$

Thus any multiset-respecting clustering of circuit points that is not a matching based clustering has a cost larger than the matching based clusterings, and the difference is at least $\frac{M\beta}{3}$. □

Lemma 2. *The formula is satisfiable if and only if there is a clustering of value at most* $\frac{kM\beta}{2} + \frac{2M}{2M+1}\alpha m$.

Proof: ($\Rightarrow$:) Consider one of the satisfying assignments of the formula. A clustering can be constructed from it as follows:

If $x_i = 1$ (respectively $x_i = 0$), cluster the points of s_i according to the true (false) matching. As every clause C_j is satisfied, fix one of the variables x_i that satisfies it. Put the clause point p_j with the nearest pair of s_i. If $x_i = 1$, then points of s_i are clustered into true matching pairs. Further, x_i appears in C_j in non-negated form, and so, by our construction, p_j is at a distance α from the midpoint of one of the true matching pairs. Thus p_j can be clustered with this pair. The cost of this cluster is

$$\begin{aligned} \text{Cost(cluster)} &= \frac{1}{2M+1}\left(M^2\beta + 2M\left(\alpha + \frac{\beta}{4}\right)\right) \\ &= \frac{2M}{2M+1}\alpha + \frac{M\beta}{2} \end{aligned}$$

The case $x_i = 0$ is analogous. Clustering all clause points in this way gives a clustering where m clusters contribute $\frac{M\beta}{2} + \frac{2M}{2M+1}\alpha$ each, and the remaining contribute $\frac{M\beta}{2}$ each, giving an overall value of $\frac{kM\beta}{2} + \frac{2M}{2M+1}\alpha m$.

($\Leftarrow$:) Suppose there is a clustering of value at most $\frac{kM\beta}{2} + \frac{2M}{2M+1}\alpha m$. By Proposition 4(2), we can assume that there is a multiset-respecting clustering $\mathcal{C}$ with this value. Let $\mathcal{C}'$ denote the restriction of $\mathcal{C}$ to circuit points. By Proposition 4(4), adding the clause points cannot decrease the cost of the clustering; thus $\text{Cost}(\mathcal{C}') \leq \text{Cost}(\mathcal{C})$. Now we prove a series of claims:

1. The restriction of $\mathcal{C}$ to circuit points, $\mathcal{C}'$, has exactly k non-empty clusters. If $\mathcal{C}'$ has fewer clusters, then there is a cluster with more than 2 points. Refine the clustering by removing one point from such a cluster and putting it in a cluster by itself. Repeat until there are exactly k clusters, to get clustering $\mathcal{C}''$ of circuit points. By Proposition 4(5), $\text{Cost}(\mathcal{C}'') \leq \text{Cost}(\mathcal{C}')$. $\mathcal{C}''$ is not matching-based (since we created singleton clusters), so by Lemma 1, it contributes a value of at least $\frac{kM\beta}{2} + \frac{M\beta}{3}$. Since

   $$\frac{kM\beta}{2} + \frac{M\beta}{3} \leq \text{Cost}(\mathcal{C}'') \leq \text{Cost}(\mathcal{C}') \leq \text{Cost}(\mathcal{C}),$$

   we have

   $$\text{Cost}(\mathcal{C}) - \left(\frac{kM\beta}{2} + \frac{2M}{2M+1}\alpha m\right) \geq \frac{M\beta}{3} - \frac{2M}{2M+1}\alpha m \geq \frac{M\beta}{6} > 0,$$

   where the second inequality follows by our choice of M. We have reached a contradiction to our assumption about $\text{Cost}(\mathcal{C})$.
2. $\mathcal{C}'$ is a matching-based clustering. That is, in $\mathcal{C}$, all circuit points are clustered into a matching based clustering.
   If not, then by the argument used above for $\mathcal{C}''$, we obtain a contradiction to our assumption about $\text{Cost}(\mathcal{C})$.

3. No cluster in $\mathcal{C}$ has more than one clause point.
   If some cluster C has two or more clause points, then let u, v be the vertices of the matching in C, and let p, q be two distinct clause points in it. By Proposition 4(4), the cost of the cluster C is at least the cost of the cluster containing just u, v, p, q. Thus using Proposition 4(1), we have

$$\text{Cost}(C) \geq \frac{1}{2M+2}\left[M^2\beta + 4M\left(\alpha + \frac{\beta}{4}\right) + \theta\right] = \frac{M\beta}{2} + \frac{4M\alpha+\theta}{2(M+1)}$$

   All other clusters have a cost of at least $M\beta/2$ each, so the overall cost is at least

$$\begin{aligned}
\text{Cost}(\mathcal{C}) &\geq (k-1)\frac{M\beta}{2} + \frac{M\beta}{2} + \frac{4M\alpha+\theta}{2(M+1)} \\
&\geq \frac{kM\beta}{2} + \frac{4M\alpha + 2(M+1)\alpha m}{2(M+1)} \qquad \text{by our choice of } \theta \\
&> \frac{kM\beta}{2} + \frac{2M}{2M+1}\alpha m \qquad \text{a contradiction.}
\end{aligned}$$

4. Each clause point is clustered with the nearest pair of circuit points, which should also be a matching pair in the matching based clustering.
   Every cluster containing a clause point has cost $\frac{M\beta}{2} + \frac{2M\alpha}{2M+1}$ if the circuit edge in the cluster is nearest the clause point, and has cost at least $\frac{M\beta}{2} + \frac{2M}{2M+1}\alpha + \frac{M\beta}{2M+1}$ otherwise.

Thus a satisfying assignment can be constructed from this clustering. □

3.3. The Details of the Layout

We now describe the layout obtained from the planar 3SAT formula F that gives us the desired instance I of the k-means problem. Let $G = (V, E)$ be the associated planar clause-variable incidence matrix. (From Definition 3, $G = (V, E_1)$. Since $G(F)$ is planar, so is G.) Note that the vertex set V of G can be partitioned into two sets: X corresponding to variable vertices, and Y corresponding to clause vertices, and G is bipartite with $E \subset X \times Y$. All vertices in Y have degree at most 3, and all vertices in X have degree at most m.

1. Let $\mathcal{E}$ be a planar combinatorial embedding of G; such an embedding can be obtained in polynomial time, and even in log space. (See for instance [1].) $\mathcal{E}$ corresponds to some plane drawing of G and specifies, for each vertex v, the cyclic ordering of the edges incident on v in this drawing.
2. Construct a related bounded-degree planar graph H and an embedding $\mathcal{E}'$ as follows: replace each vertex $v \in X$ by a cycle C_v on m vertices, $v_1, v_2, \ldots, v_m$. Reroute the $d(v)$ edges incident on v in G to the first $d(v)$ of these vertices, in the same order as dictated by $\mathcal{E}$. It is straightforward to see that H is planar, and its embedding $\mathcal{E}'$ can be easily obtained from $\mathcal{E}$. The maximum degree of any vertex in H is 3. The vertex set of H is the disjoint union of X' and Y, where $X' = X \times [m]$.

3. Consider a plane drawing of H where vertices are embedded at points on an integer grid, and edges are embedded as rectilinear paths. Such a drawing can be obtained in polynomial time [15, 21], and even in logarithmic space [1].
4. Inflate the grid by a factor of $b \geq 14$.
   This ensures, in particular, that every vertex or bend point u is at the centre of a big box B_u of size $b \times b$, and a small box S_u of size 6×6. The big boxes for different grid points have disjoint interiors, and thus contain no other vertex or bend point even on their boundaries.
   Consider an edge connecting vertex $[x, k] \in X'$ with vertex $y \in Y$. Replace it by a pair of parallel rectilinear paths separated by two grid squares. At the y end, join up these paths along the boundary of S_y. At the $[x, k]$ end, splice them along with the edges to $[x, k-1]$ and $[x, k+1]$ to form a continuous path. See Figure 1. Note that some additional rectilinear bends might be required at the $[x, k]$ end, (see, for example, $(x, 4)$ in Figure 1).

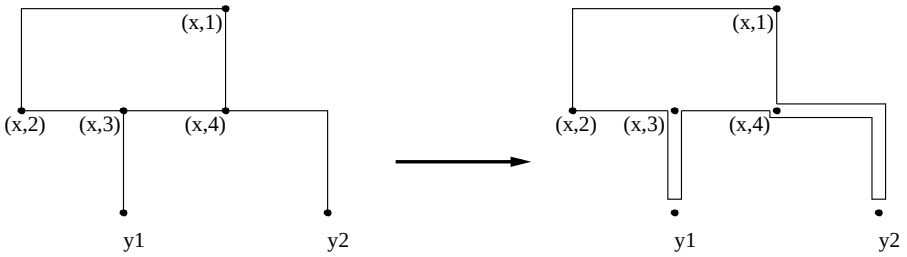

Figure 1: Creating circuits for variables

   For each vertex $x \in X$ (and hence for each variable in F), this process distorts the cycle C_x in H into a circuit t. Let t_i denote the circuit corresponding to variable x_i. Since t_i is a rectilinear circuit on a grid, it is of even length.
5. Each clause point y_j is now moved to the center of one of the grid squares touching it, the one that is to the North-West. Extend the three circuits "incident" to the clause point, if necessary, so that all incident circuits are at a Euclidean distance of precisely $\frac{5}{2}$ times the grid length from the moved clause point. See the layout and the modification in Figure 2.

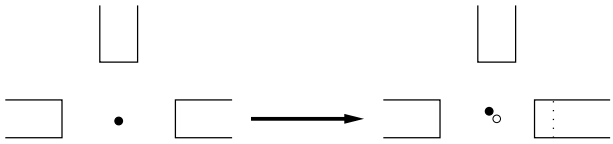

Figure 2: Repositioning clause points

6. For each circuit t_i, arbitrarily fix one of its perfect matchings as the true matching, and the other as the false matching.

Let clause C_j contain variable x_i positively (negatively, respectively). If in the layout so far, y_j is nearest a true (false, resp.) edge of t_i, then nothing needs to be done. If, however, the edge of t_i nearest y_j is a false (true, resp.) edge, further deform t_i in the area within B_{y_j} but outside S_{y_j}. Replace a sub-path of length two (on each parallel path) by a path of length three, with the vertices laid out on a regular semi-hexagon and hence at distance one from their neighbours on the circuit. Change the true/false matchings within B_{y_j} to be consistent with the labelling outside. This makes the edge nearest y_j a true edge if it was false earlier, and vice versa. The overall length of the circuit remains even. See Figure 3.

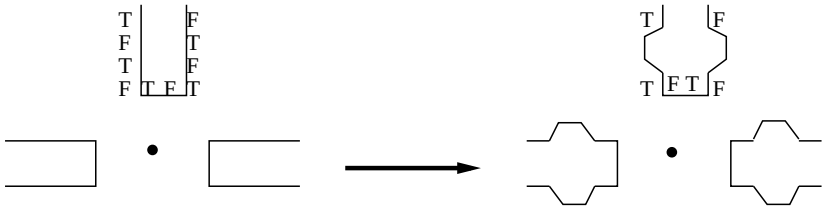

Figure 3: Adjusting the parity of circuits relative to clause points

Since the grid was inflated sufficiently, these distortions do not affect other vertices / bends.
After doing this distortion wherever needed, the resulting circuit for a variable is the required circuit s_i.

Figure 4 gives the complete layout for a small planar 3SAT instance.

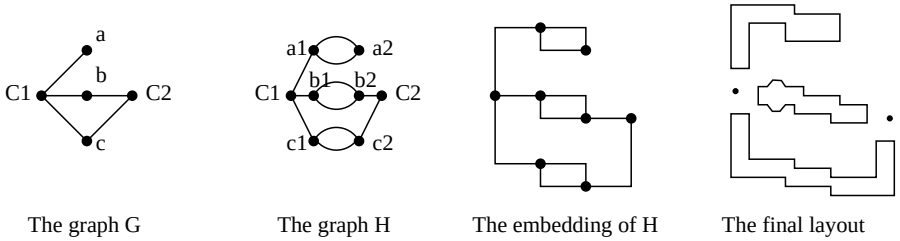

Figure 4: The layout for $F = (a \vee b \vee c) \wedge (\overline{b} \vee c)$

Let the squared unit length of the grid be β. Then $\alpha = (\frac{5}{2})^2\beta = 6.25\beta$. Any two clause points are separated either vertically or horizontally by b grid lengths, so the distance between them is at least $\theta = b^2\beta$.

Figure 5 shows the box B_u for a clause point u, and within this box the smallest distances are demonstrated. The nearest vertices are A_3 and A_4 (and the corresponding vertices on the other two circuits as well). The distances satisfy the following:

point pair	distance	point pair	distance
		u, A_2	$\alpha + 6\beta + \beta/4$
A_i, A_{i+1}	β	u, A_3	$\alpha + \beta/4$
A_1, A_3	3β	u, A	α
A_2, A_4	2β	u, A_4	$\alpha + \beta/4$
A_3, A_5	4β	u, A_5	$\alpha + 2\beta + \beta/4$

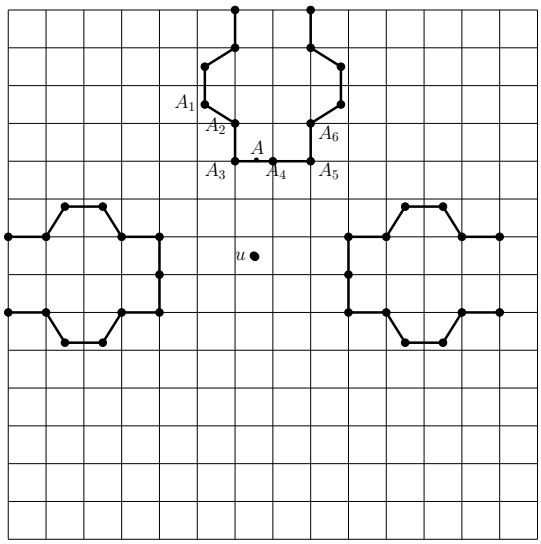

Figure 5: The distances α, β shown inside B_u for a clause point u.

It is straightforward to see that with $M = 38m$, $b = 28m$, all the required conditions on the parameters are satisfied.

3.4. Dealing with the irrational coordinates

In the last step of the layout, where we replace certain sub-paths of length 2 by sub-paths of length 3, the numerators of the point coordinates become irrational. Essentially, we introduce multiples of $\sqrt{3}$ in the numerator. However, there is a gap in Lemma 2 between the k-means clustering costs corresponding to satisfiable and unsatisfiable instances. So we may "round" these irrational points to sufficiently close rational points.

Lemma 2 shows that the optimal k-means clustering cost is at most $\mu = \frac{kM\beta}{2} + \frac{2M}{2M+1}\alpha m$ in the case where the original planar 3-CNF formula is satisfiable; a quick glance at the proof shows that if the original formula is not satisfiable, the optimal k-means clustering cost is at least $\mu + \lambda$, where

$$\lambda = \min\left(\frac{M\beta}{6}, \frac{2M\alpha}{M+1}, \frac{M\beta}{2M+1}\right).$$

Let $\varepsilon = \frac{\lambda/2}{\mu+\lambda} < \frac{\lambda/2}{\mu}$.

Let d_{min} denote the smallest Euclidean distance between two different locations in the layout. We consider a simplistic rounding: for a location with coordinates (x, y) where, say, x is irrational (only one coordinate is irrational in the construction so far), move the location to (x', y) so that x' is rational, and $|x' - x| < \frac{\varepsilon d_{min}}{8}$. Observe that it is possible to find such an x' in polynomial time ([18], Theorem 1.4.7). Now if p and q denote two points in the original layout, and p' and q' denote the corresponding points in the rounded layout, then

$$d(p', q') < d(p, q) + \frac{\varepsilon d_{min}}{4} \leq d(p, q) \left[1 + \frac{\varepsilon}{4}\right]$$

(recall, $d(u, v)$ is the Euclidean distance between u and v) and so

$$\text{dist}(p', q') < \text{dist}(p, q) \left[1 + \frac{\varepsilon}{2} + \frac{\varepsilon^2}{8}\right] \leq \text{dist}(p, q) \left[1 + \varepsilon\right]$$

Similarly, we can see that $\text{dist}(p', q') > \text{dist}(p, q) \left[1 - \varepsilon\right]$. Thus,

$$(1 - \varepsilon)\text{dist}(p, q) < \text{dist}(p', q') < (1 + \varepsilon)\text{dist}(p, q).$$

Now, Proposition 4 (1) implies that for any clustering of the points, the ratio of the cost in the rounded layout to the cost in the original layout is strictly greater than $(1 - \varepsilon)$ and strictly less than $(1 + \varepsilon)$.

Thus, the optimal k-means clustering cost in the rounded layout is strictly less than $(1+\varepsilon)\mu \leq \mu+\lambda/2$ when the input formula is satisfiable, and is strictly greater than $(1-\varepsilon)(\mu+\lambda) \geq \mu+\lambda/2$ when the input formula is not satisfiable.

4. Discussion

We have shown that the k-means clustering problem remains NP-complete even in two dimensions, when the number of centers k is part of the input. The NP-hardness of this problem has been independently observed by Andrea Vattani, [22].

There are still some unsettled issues regarding this hardness. An obvious question is whether there are natural parameters associated with planar k-means instances such that when these parameters are resricted in some way, the problem becomes tractable.

- One possible choice of the parameter is k, the number of centers itself. It is known that for planar k-means with constant values of k, there is a polynomial-time algorithm due to [12]. Our reduction places no bounds on the value of k; it is unrestricted (and in particular, can be as large as $\theta(n)$). A natural question to ask is where is the hardness threshold; at what values of k does the planar k-means problem become NP-hard. For instance, for $k \in O(\log n)$, the algorithm by [12] runs in quasi-polynomial time, and thus is unlikely to be NP-hard. Our reduction shows hardness for a particular choice of ϵ. Is it hard for $k = n^\epsilon$ for every choice of $\epsilon \in (0, 1)$? It has been pointed out by Vattani ([22]) that this is indeed the case.

- Another possible parameter to examine is the ratio of the maximum distance between points to the minimum distance. In an instance generated in our reduction, this ratio is infinity, as there are points with distance 0. A small perturbation of the points will make this ratio finite, but still unbounded. (On the other hand, it will be polynomial in n.) If the ratio is known a priori to be, say, linear in n, does it make the problem easier?

  However, if we consider only different locations, then the ratio even in our reduction is bounded by a polynomial in n, as the grid itself is of size polynomial in n. It is not clear if linear ratio is possible preserving hardness.

Regarding approximability also, the picture concerning planar k-means is far from clear. The algorithm of [6] (a variant of Lloyd's algorithm), while providing approximation guarantees for general k-means, does not provide any better guarantees on planar instances. (Although the lower bound example constructed in [6] is for high dimensions, analogous planar instances can also be constructed.) However, its behaviour on planar instances is not fully understood. In particular, it is entirely possible that for any planar instance of k-means, the algorithm of [6] gives an $O(1)$-approximation with high probability. Even if this is not true, it may still hold for most planar instances. Settling this either way would be of some interest.

The most important open question is to determine whether there is a PTAS for the planar k-means problem. Note that for a very similar problem, the *planar k-median problem*, a PTAS is known to exist ([3]). This is despite the fact that unlike the 1-mean, the 1-median does not have a closed form solution.

Acknowledgement

The authors would like to thank Amit Deshpande for useful discussions concerning the behaviour of [6] on planar instances.

References

[1] E. Allender, S. Datta, and S. Roy. The directed planar reachability problem. In *Proc. 25th Foundations of Software Technology and Theoretical Computer Science Conference*, volume 3821, pages 238–249. LNCS, 2005.

[2] D. Aloise, A. Deshpande, P. Hansen, and P. Popat. NP-hardness of euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009.

[3] S. Arora. Polynomial time approximation schemes for euclidean tsp and other geometric problems. In *FOCS '96: Proceedings of the 37th Annual Symposium on Foundations of Computer Science*, page 2, 1996.

[4] D. Arthur and S. Vassilvitskii. How slow is the k-means method? In *Proc. Symp. on Comput. Geom.*, 2006.

[5] D. Arthur and S. Vassilvitskii. Worst-case and smoothed analysis of the icp algorithm, with an application to the k-means method. In *Proc. IEEE Symp. Foundations of Computer Science*, 2006.

[6] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proc. ACM-SIAM Symp. Discrete Algorithms*, 2007.

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

[8] F. de la Vega, M. Karpinski, and C. Kenyon. Approximation schemes for clustering problems. In *Proc. ACM Symp. Theory of Computing*, pages 50–58, 2003.

[9] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Machine Learning*, 56:9–33, 2004.

[10] M. Gibson, G. Kanade, E. Krohn, I. Pirwani, and K. Varadarajan. On clustering to minimize the sum of radii. In *Proc. ACM-SIAM Symp. Discrete Algorithms*, 2008.

[11] S. Har-Peled and B. Sadri. How fast is the k-means method? In *Proc. ACM-SIAM Symp. Discrete Algorithms*, pages 877–885, 2005.

[12] M. Inaba, N. Katoh, and H. Imai. Applications of weighted voronoi diagrams and randomization to variance-based clustering. In *Proc. Annual Symp. on Comput. Geom.*, pages 332–339, 1994.

[13] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu. A local search approximation algorithm for k-means clustering. *Comput. Geom.*, 28:89–112, 2004.

[14] A. Kumar, Y. Sabharwal, and S. Sen. A simple linear time $(1+\epsilon)$ approximation algorithm for k-means clustering in any dimensions. In *Proc. IEEE Symp. Foundations of Computer Science*, pages 454–462, 2004.

[15] C.E. Leiserson. Area-efficient graph layouts (for VLSI). In *Proc. 21st Ann. IEEE Symp. Foundations of Computer Science*, pages 270–281, 1980.

[16] D. Lichtenstein. Planar formulae and their uses. *SIAM J. Comput.*, 11:329–343, 1982.

[17] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–136, 1982.

[18] L. Lovász. An algorithmic theory of numbers, graphs, and convexity. *CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM*, 1986.

[19] N. Megiddo and K. Supowit. On the complexity of some common geometric location problems. *SIAM J. Comput.*, 13:182–196, 1984.

[20] R. Ostrovsky, Y. Rabani, L. Schulman, and C. Swamy. The effectiveness of lloyd-type methods for the k-means problem. In *Proc. IEEE Symp. Foundations of Computer Science*, 2006.

[21] L. G. Valiant. Universality considerations in vlsi circuits. *IEEE Transactions on Computers*, 30:135–140, 1981.

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