Geometric clustering in normed planes *

Pedro Martín[1] and Diego Yáñez[2]

1 **Departamento de Matemáticas, Universidad de Extremadura**
pjimenez@unex.es
2 **Departamento de Matemáticas, Universidad de Extremadura**
dyanez@unex.es

Abstract

Given two sets of points A and B in a normed plane, we prove that there are two linearly separable sets A' and B' such that $\mathrm{diam}(A') \leq \mathrm{diam}(A)$, $\mathrm{diam}(B') \leq \mathrm{diam}(B)$, and $A' \cup B' = A \cup B$. As a consequence, some Euclidean clustering algorithms are adapted to normed planes.

1 Introduction and notation

We denote by $\mathbb{E}^2$ the Euclidean plane, and by $\mathbb{M}^2$ a *normed plane*, namely, $\mathbb{R}^2$ endowed with a norm $\|\cdot\|$. We call $B(x, r)$ the *ball with center* $x \in \mathbb{M}^2$ *and radius* $r > 0$, and $S(x, r)$ the *sphere* of $B(x, r)$. We use the usual abbreviations $\mathrm{diam}(A)$ and $\mathrm{conv}(A)$ for the *diameter* and the *convex hull* of a set A, $\overline{ab}$ for the *line segment* connecting two points $a, b \in \mathbb{M}^2$, and $\langle a, b \rangle$ for its affine hull.

We say that two sets of points in $\mathbb{M}^2$ are *linearly separable* (for short, *separable*) if there exists a line L such that each set is situated in a different closed half-plane defined by L. In Section 2, our Theorem 2.3 extends the following result ([4]) to any normed plane.

▶ **Theorem 1.1.** *Let A and B be two finite sets in $\mathbb{E}^2$. Then, there are two separable sets A' and B' such that* $\mathrm{diam}(A') \leq \mathrm{diam}(A)$, $\mathrm{diam}(B') \leq \mathrm{diam}(B)$, *and* $A' \cup B' = A \cup B$.

Given a set S of n points in the plane, a *cluster* is any non-empty subset of S, and a k-*clustering* is a set of k clusters such that each point of S belongs to some cluster. In Section 3, we apply Theorem 2.3 in order to solve some k-clustering problems in any normed plane.

2 Linear separability of clusters

In the rest of this section we work in $\mathbb{M}^2$ and our objective is to prove the statement of Theorem 1.1 in this context. Without loss of generality, we assume that $\mathrm{diam}(A) \geq \mathrm{diam}(B)$. Let us denote $\{u_1, u_2, \ldots, u_{2t}\}$ the clockwise sequence of points where the boundaries of $\mathrm{conv}(A)$ and $\mathrm{conv}(B)$ cross (Figure 1). $\mathrm{conv}(A) \setminus \mathrm{conv}(B)$ and $\mathrm{conv}(B) \setminus \mathrm{conv}(A)$ consist of two interlacing sequences of polygons $\{A_1, A_2, \ldots, A_t\}$ and $\{B_1, B_2, \ldots, B_t\}$ such that (for convenience, $u_{2t+1} := u_1$ and $A_{t+1} := A_1$): A_i touches B_i at u_{2i}; B_i touches A_{i+1} at u_{2i+1}; the vertices of any A_i belong either to $A \setminus B$ or to $\mathrm{conv}(A) \cap \mathrm{conv}(B)$; the vertices of any B_j belong either to $B \setminus A$ or to $\mathrm{conv}(A) \cap \mathrm{conv}(B)$. We say that (A_i, B_j) is a *bad pair* if $\mathrm{diam}(A_i \cup B_j) > \mathrm{diam}(A)$. In such a case, A_i is a *bad set* and B_j is its *bad partner*, and vice versa. If $\|a_i - b_j\| > \mathrm{diam}(A)$ for some $a_i \in A_i$ and $b_j \in B_j$, then both a_i and b_j are *bad points*, a_i is a *bad partner* of b_j (and vice versa), and the segment $\overline{a_i b_j}$ is a *bad segment*.

▶ **Lemma 2.1.** *Let (A_i, B_j) and $(A_{i'}, B_{j'})$ be two bad pairs such that $A_i \neq A_{i'}$ and $B_j \neq B_{j'}$. Let us choose $a_i \in A_i, b_j \in B_j, a_{i'} \in A_{i'}, b_{j'} \in B_{j'}$ such that $\overline{a_i b_j}$ and $\overline{a_{i'} b_{j'}}$ are bad segments.*

* Partially supported by Junta de Extremadura grant GR15055 (partially financed with FEDER)

34th European Workshop on Computational Geometry, Berlin, Germany, March 21–23, 2018.

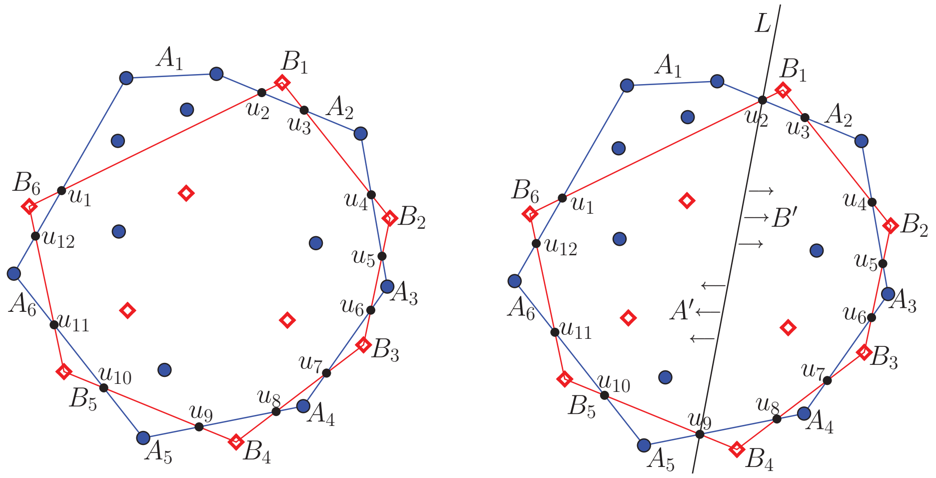

Figure 1 A (blue points) and B (red points) are not separable (left). $A \cup B$ can be split by L into new subsets A' and B' without increase of the Euclidean diameters (right).

If these bad segments do not cross, then $A_i, A_{i'}, B_{j'}, B_j$ (disregarding symmetric variations) is the sequence clockwise of these polygons and there is not any bad set from A between $B_{j'}$ and B_j.

Proof. Let us assume that $A_i, A_{i'}, B_j, B_{j'}, a_i, a_{i'}, b_j$, and $b_{j'}$ satisfy the conditions of the Lemma. All of them must be situated around $\text{conv}(A \cap B)$. If $\overline{a_i b_j} \cap \overline{a_{i'} b_{j'}} = \emptyset$, there are two cases (disregarding symmetric variations) for the relative positions of the polygons (and points):

Case 1: *$A_i, B_{j'}, A_{i'}, B_j$ is the clockwise sequence of the polygons.* Since the sum of the diagonals of the quadrangle $a_i, b_{j'}, a_{i'}, b_j$ is larger than the sum of two opposite sides, we get a contradiction:

$$\text{diam}(A) + \text{diam}(B) \geq \|a_i - a_{i'}\| + \|b_j - b_{j'}\| \geq \|a_i - b_j\| + \|a_{i'} - b_{j'}\| > 2\,\text{diam}(A).$$

Case 2: *$A_i, A_{i'}, B_{j'}, B_j$ is the clockwise sequence of the polygons.* Let us assume that there exists a bad point $a_m \in A_m$ for some m, such that $B_{j'}, A_m, B_j$ is the clockwise sequence. Let b_k be a bad partner of a_m for some k. The half-lines starting in a_m and connecting a_m with a_i and with $a_{i'}$, and the lines $\langle a_m, b_j \rangle$ and $\langle a_m, b_{j'} \rangle$, divide the plane into six zones (see Figure 2). If b_k is situated in the shaded zone in Figure 2, then $\|a_m - b_k\| \leq \text{diam}(B)$ and $\overline{a_m b_k}$ is not a bad segment. If b_k belongs to any other zone, it is possible to consider a quadrangle whose vertices are situated in clockwise order like in Case 1, and we get a contradiction. ◀

▶ Remark 1. In the Euclidean subcase, every two bad segments from disjoint bad pairs (A_i, B_j) and $(A_{i'}, B_{j'})$ cross ([4]). The property that the longest side of every obtuse triangle is opposite to the obtuse angle is used in the proof. Nevertheless, this property is not true for any normed plane, and there exist bad segments that do not cross.

Before splitting the sets A and B, we group all the adjacent bad subsets A_i from the cluster A. Thus, we define a group of bad subsets from A to be a maximal cyclic subsequence of bad subsets A_i. (Intervening subsets B_j of the other cluster must not be bad). The same

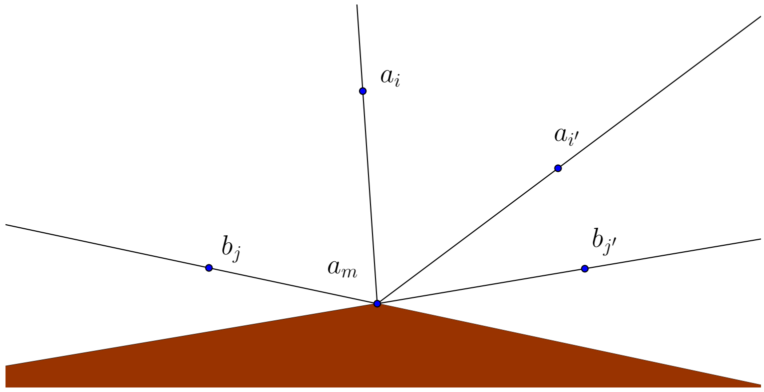

Figure 2 If (a_i, b_j) and $(a_{i'}, b_{j'})$ are bad partners, then the shaded zone cannot contain a bad partner of $a_m \in A_m$

is made with cluster B. These maximal cyclic groups are noted by $\bar{A}_1, \bar{A}_2, \ldots, \bar{A}_p$ and $\bar{B}_1, \bar{B}_2, \ldots, \bar{B}_q$.

We say that $(\bar{A}_i, \bar{B}_j)$ is a *bad pair of groups* if there exists a bad segment from $\bar{A}_i$ to $\bar{B}_j$. Two pairs of sets (A_i, B_j) and $(A_{i'}, B_{j'})$ cross if there exist two (one from every pair) bad segments that cross. Similarly, $(\bar{A}_i, \bar{B}_j)$ and $(\bar{A}_{i'}, \bar{B}_{j'})$ cross if there exist two (one from every pair) bad segments that cross.

The structure of the rest of the section is similar to that presented by [4], but the proofs are different due to Remark 1. Some of the these proofs are omitted in this extended abstract.

▶ **Lemma 2.2.** *The following holds:*

1. *Let (A_i, B_j) and $(A_{i'}, B_{j'})$ be two bad pairs such that $A_i \neq A_{i'}$ and $B_j \neq B_{j'}$. If A_i and $A_{i'}$ belong to a group $\bar{A}_n$ for some n, then B_j and $B_{j'}$ belong to a group $\bar{B}_t$ for some t.*
2. *The number of maximal cyclic groups for A and for B is the same.*
3. *Let $(\bar{A}_i, \bar{B}_j)$ and $(\bar{A}_{i'}, \bar{B}_{j'})$ be two bad pairs of groups such that $\bar{A}_i \neq \bar{A}_{i'}$ and $\bar{B}_j \neq \bar{B}_{j'}$. Then $(\bar{A}_i, \bar{B}_j)$ and $(\bar{A}_{i'}, \bar{B}_{j'})$ cross.*

Proof. Statement 2 is a consequence of 1. In order to prove 1, let us assume that (A_i, B_j) and $(A_{i'}, B_{j'})$ are two disjoint bad pairs such that $A_i, A_{i'} \in \bar{A}_n$ for some n. If B_j and $B_{j'}$ belong to different groups, there is a bad pair (A_m, B_k) for some m and k such that A_m separates B_j from $B_{j'}$. (A_i, B_j) and $(A_{i'}, B_{j'})$ must cross (Lemma 2.1), and since A_i and $A_{i'}$ belong to the same group, only one of the pairs (not both) and (A_m, B_k) cross. For simplicity, let us assume that (A_i, B_j) and (A_m, B_k) cross. There exist $a_m \in A_m, b_k \in B_k, a_{i'} \in A_{i'}, b_{j'} \in B_{j'}$ that would be situated in an impossible clockwise sequence $a_m, b_k, a_{i'}, b_{j'}$ (similar to Case 1 in Lemma 2.1), and we get a contradiction.

Let us see Statement 3. Let $(\bar{A}_i, \bar{B}_j)$ and $(\bar{A}_{i'}, \bar{B}_{j'})$ be two bad pairs of groups such that $\bar{A}_i \neq \bar{A}_{i'}$ and $\bar{B}_j \neq \bar{B}_{j'}$. The clockwise order cannot be $\bar{A}_i, \bar{B}_{j'}, \bar{A}_{i'}, \bar{B}_j$ (due to the arguments used in Case 1 of Lemma 2.1); and neither $\bar{A}_i, \bar{A}_{i'}, \bar{B}_{j'}, \bar{B}_j$, because then $\bar{B}_{j'}$ and $\bar{B}_j$ cannot be separated by a bad polygon A_m (Lemma 2.1, Case 2). Therefore, the clockwise order must be $\bar{A}_i, \bar{A}_{i'}, \bar{B}_j, \bar{B}_{j'}$, and 3 holds.

◀

The groups from A and B are interlacing, and Statement 3 of Lemma 2.2 implies that there exist a complete matching among the groups, and the number of groups from each cluster has to be odd.

Let A_i be the last bad set of a group (in clockwise order), and let $B_{j'}$ be the last bad partner of A_i. Let B_j be the first bad set after A_i, and let $A_{i'}$ be the first bad partner of B_j. We choose the separating line L to go through the point u_{2j} before B_j and the point $u_{2j'+1}$ after $B_{j'}$ (see Figure 1). We define B' to be the points in $A \cup B$ lying on the same side of L as B_j and $B_{j'}$, and A' as the remaining points.

▶ **Theorem 2.3.** *Let A and B be two finite sets in $\mathbb{M}^2$. Then, there are two linearly separable sets A' and B' such that* $\mathrm{diam}(A') \leq \mathrm{diam}(A)$, $\mathrm{diam}(B') \leq \mathrm{diam}(B)$, $A' \cup B' = A \cup B$, *and*

$$\mathrm{perimeter}(\mathrm{conv}(A)) + \mathrm{perimeter}(\mathrm{conv}(B)) \geq \mathrm{perimeter}(\mathrm{conv}(A')) + \mathrm{perimeter}(\mathrm{conv}(B')).$$

Proof. We consider $A_i, B_j, A_{i'}, B_{j'}, L, A'$ and B' defined above. Since L cuts all bad pairs, $\mathrm{diam}(A') \leq \mathrm{diam}(A)$. In order to prove $\mathrm{diam}(B') \leq \mathrm{diam}(B)$, let us consider $a, b \in B'$. If $a, b \in B$ there is nothing to prove. In any other case, let us assume that $\|a - b\| > \mathrm{diam}(B)$. Let us choose $a_i \in A_i$, $b_j \in B_j$, $a_{i'} \in A_{i'}$, $b_{j'} \in B_{j'}$ such that $(a_i, b_{j'})$ and $(a_{i'}, b_j)$ are bad pairs. There are three possible cases.

Case 1: $a \in \mathrm{conv}(A) \setminus \mathrm{conv}(B)$ and $b \in \mathrm{conv}(B) \setminus \mathrm{conv}(A)$. The points $\{b_j, a, b, b_{j'}, a_{i'}, a_i\}$ are situated around $\mathrm{conv}(A) \cap \mathrm{conv}(B)$ and it is possible to consider a clockwise order. If $\{a, b\}$ is the clockwise order of these two points, we observe the quadrangle with vertices (clockwise) $\{b_j, a, b, a_{i'}\}$ and the following contradiction holds:

$$\mathrm{diam}(A) + \mathrm{diam}(B) \geq \|a - a_{i'}\| + \|b - b_j\| \geq \|b_j - a_{i'}\| + \|a - b\| > \mathrm{diam}(A) + \mathrm{diam}(B).$$

If the clockwise order is $\{b, a\}$, we obtain a similar contradiction on the quadrangle with vertices (clockwise order) $\{a_i, b, a, b_{j'}\}$.

Case 2: $a, b \in \mathrm{conv}(A) \setminus \mathrm{conv}(B)$. Case 1 implies that $\|b - b'\| \leq \mathrm{diam}(B)$ for every $b' \in (\mathrm{conv}(B) \setminus \mathrm{conv}(A)) \cap B'$. If $\{a, b\}$ is the clockwise order of these two vertices, applying the above arguments to the quadrangle $\{b_j, a, b, a_{i'}\}$:

$$\mathrm{diam}(A) + \mathrm{diam}(B) \geq \|a - a_{i'}\| + \|b - b_j\| \geq \|b_j - a_{i'}\| + \|a - b\| > \mathrm{diam}(A) + \mathrm{diam}(B),$$

which is again a contradiction. If the order is $\{b, a\}$, we use the quadrangle $\{b_j, b, a, a_{i'}\}$.

Case 3: $a \in \mathrm{conv}(A) \setminus \mathrm{conv}(B)$ and $b \in \mathrm{conv}(A) \cap \mathrm{conv}(B)$. Since the distance from a is maximized at some vertex of $\mathrm{conv}(A) \cap \mathrm{conv}(B) \cap \mathrm{conv}(B')$, we may assume that b is one of these vertices and apply an analysis similar to Case 1 or to Case 2.

The proof in [4] for the perimeter inequality is valid for $\mathbb{M}^2$. ◀

3 Some applications to clustering problems

From now on, S is a set of n points in $\mathbb{M}^2$. We assume that in our computation model an oracle answers the required questions about the unit ball of $\mathbb{M}^2$ (see Section 3.3 of [6]).

3.1 2-clustering problem: minimize the maximum diameter.

Given a metric, the 2-*clustering problem of minimizing the maximum diameter* asks about how to split S into two sets minimizing the maximum diameter. Avis [1] solves the problem in $\mathbb{E}^2$ looking for two separable sets with the following algorithm ($O(n^2 \log^2 n)$ time): sort the distances d_i between the points of S into increasing order ($O(n^2 \log n)$ time); locate the minimum d_i that admits a *stabbing line*[1] (using [5] for the stabbing line) by a binary search. We obtain the following as a consequence of Theorem 2.3.

[1] A *stabbing line* for a set of segments is a line that intersects every segment of the set.

▶ **Corollary 3.1.** *Given a set of n points in $\mathbb{M}^2$, the 2-clustering problem of minimizing the maximum diameter can be solved in $O(n^2 \log^2 n)$ time using the algorithm presented by Avis.*

The similar approach of Asano et al. ([2], that reduces the cost of Avis' approach to $O(n \log n)$ time using a maximum spanning tree) could be used as well, but as far as we know, an efficient method to build a maximum spanning tree for any normed plane is not known.

3.2 2-clustering problem: constraints over the diameters

Given two fixed numbers $d_1 \geq d_2 > 0$, Hershberger and Suri ([8]) solve in $O(n \log n)$ time the problem of dividing S into two sets S_1 and S_2, such that $\mathrm{diam}(S_1) \leq d_1$ and $\mathrm{diam}(S_2) \leq d_2$ in $\mathbb{E}^2$. They use the fact that if $\|a - b\| \geq d_1$, then $B(a, d_2) \cap B(b, d_1)$ can always be split into two subsets whose diameters are at most d_1 and d_2, respectively. Nevertheless, the following example shows that this cannot be extended to $\mathbb{M}^2$. Let us consider $a = (0, 0)$, $b = (-9.81, 6.24)$, and the strictly convex norm whose unit sphere is bounded by the two arcs of circles with center at $(0, 10)$ and in $(0, -10)$, respectively, and radius $5\sqrt{13}$ (see Figure 3). Let $\{r = (-9.39, r_2), \quad s = (-8.24, s_2)\} \in S(a, 1)$ and $\{p, q\} = S(a, 1) \cap S(b, 1.1)$, such that $r_2 > 0$, $s_2 > 0$, and $\{p, r, s, q\}$ is the clockwise order on $S(a, 1)$. Then, $\|a - b\| \geq 1.1$, $\min\{\|s - p\|, \|r - q\|, \|p - q\|\} > 1.1$ and $\min\{\|r - p\|, \|s - q\|\} > 1$, and $S = \{p, q, r, s\} \in B(a, 1) \cap B(b, 1.1)$ cannot be divided into two subsets whose diameters are at most 1.1 and 1, respectively.

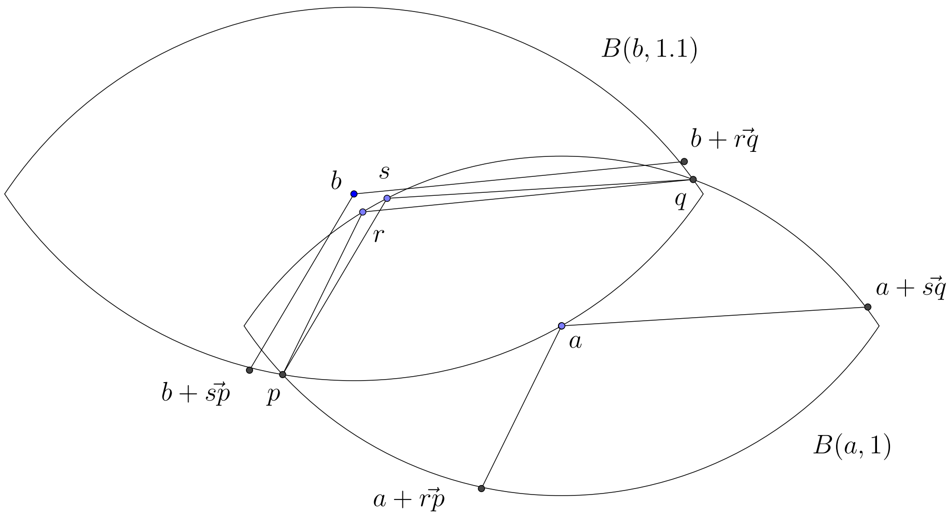

Figure 3 $S = \{p, q, r, s\}$ cannot be divided into two subsets with diameters less than or equal to 1.1 and 1, respectively.

Theorem 2.3 can help to solve this problem in any normed plane as follows. Build the graph (S, E_{d_1}) with the points of S and the set of edges E_{d_1} connecting two points of S at distance more than d_1 (in $O(n^2 \log n)$ time). Check if E_{d_1} has a stabbing line (in $O(n \log n)$ time with the algorithm presented in [5]). If the stabbing line does not exist, there is no solution (Theorem 2.3). If some stabbing lines exist, check if one of them split S into two subsets with the required diameters.

3.3 k-clustering problems

Let us consider the *k-clustering problem of minimizing $\mathcal{F}$ to the diameters* (equivalently, *to the radii*), where $\mathcal{F}$ is a monotone increasing function $\mathcal{F} : \mathbb{R}^k \to \mathbb{R}$ that is applied to the diameters (equivalently, to the radii) of the clusters. For instance, $\mathcal{F}$ can be the *maximum*, the *sum*, or the *sum of squares* of the diameters (or the radii). Capoyleas, Rote and Woeginger (see Lemma 8 and Theorem 9 in [4] for details) design an algorithm that solves these geometric k-clustering problems in polynomial time. Using Theorem 2.3 and a result of Banasiak [3] describing precisely the intersection of two balls, we prove the following statements. The proofs are omitted in this extended abstract.

▶ **Theorem 3.2.** *Let S be a set of n points in $\mathbb{M}^2$. Consider the k-clustering problem of minimizing a monotone increasing function $\mathcal{F} : \mathbb{R}^k \to \mathbb{R}$ that is applied to the diameters or to the radii of k subsets of S. Then there is an optimal k-clustering such that each pair of clusters is linearly separable. A solution can be obtained by the algorithm presented by Capoyleas-Rote-Woeginger, and it takes polynomial time for the case of the diameters.*

3.4 3-clustering problems: minimize the maximum diameter

▶ **Theorem 3.3.** *Given a set of n points in $\mathbb{M}^2$ and $d > 0$, we can determine in $O(n^3 \log^2 n)$ time whether there is a partition of S into sets A, B, C with diameters at most d, and construct in $O(n^3 \log^3 n)$ time a 3-partition of S such that the largest of the three diameters is as small as possible.*

Proof. (Scheme) A specific approach by Hagauer and Rote proves this result in $\mathbb{E}^2$. For the first statement, the authors use some lemmas (from Lemma 3 to Lemma 6 in [7]) and Theorem 1.1. Theorem 2.3 extends Theorem 1.1, and we prove results similar to the rest of lemmas in [7] for any normed plane using the notion of Birkhoff orthogonality instead of the Euclidean one. Regarding the complexity of the algorithm, we can justify that the data structure introduced by Hershberger and Suri ([8]) is usable in the same way that in $\mathbb{E}^2$. Finally, a binary search on the $\binom{n}{2}$ distances occurring in S solves the optimization problem. ◀

References

1 D. Avis. Diameter partitioning. *Discrete Comput. Geom.*, 1:265-276, 1986.

2 T. Asano, B. Bhattacharya, M. Keil, and F. Yao. Clustering algorithms based on minimum and maximum spanning trees. *Proc. 4th ACM Symposium on Computational Geometry*, 252-257, 1988.

3 J. Banasiak. Some contributions to the geometry of normed linear spaces. *Math. Nachr.*, 139:175-184, 1988.

4 V. Capoyleas, G. Rote, and G. Woeginger. Geometric clustering. *J. Algorithms*, 12:341-356, 1991.

5 H. Edelsbrunner, H.A. Mauer, F.P. Preparata, A.L. Rosenberg, E. Welzl, and D. Wood. Stabbing line segments. *BIT*, 22:274-281, 1982.

6 P. Gritzmann, and V.L. Klee. On the complexity of some basic problems in computational convexity: I. Containment problems. *Discrete Math.*, 136:129-174, 1994.

7 J. Hagauer, and G. Rote. Three-clustering of points in the plane. *Comput. Geom.*, 8:87-95, 1997.

8 J. Hershberger, and S. Suri. Finding tailored partitions. *J. Algorithms*, 12:431–463, 1991.