Clustering Points with Line Segments under the Hausdorff Distance is NP-hard

Hugo Akitaya*, Majid Mirzanezhad†, Maarten Löffler‡, Carola Wenk§

1 Introduction

Given a set of points P in $\mathbb{R}^d$ and a real number $\varepsilon > 0$, the aim is to compute the minimum number of line segments such that the undirected Hausdorff distance between the union of these segments and P is at most ε. This problem arises in various fields, such as computational geometry, GIS, shape reconstruction, and clustering.

In GIS, efficient representation of spatial data is crucial. Covering point data with linear feature segments under proximity metric allows for simplified representations of geographic features such as roads and boundaries [10]. This simplification can improve storage efficiency and computational performance in spatial queries. Shape reconstruction and pattern recognition from point clouds are fundamental problems in computer graphics and computer vision. Approximating shapes using other alternative shapes within a given error tolerance is necessary for rendering and modeling [6]. Efficient algorithms for this problem enable the reconstruction of surfaces from unorganized point sets. In clustering, grouping points based on proximity and fitting geometric objects to these groups is a common approach. The problem of covering points with segments under Hausdorff distance constraints relates to line clustering and can be used in pattern recognition and data analysis [7].

Although the problem of interest is closely related to the geometric set cover (GSC) problem where ranges are assumed to be cylindrical shapes of radii ε around the segments, the difference between the two problems remains elusive. The geometric set cover problem involves covering points with the minimum number of geometric objects (ranges) from a predefined family, such as disks, rectangles, or lines, but such shapes are not required to be entirely in the union of ε-radius balls centered at the points. This can then be seen as a version of our problem with directed Hausdorff distance. It is a well-studied NP-hard problem [3]. Approximation algorithms have been developed for various geometric set cover problems, often leveraging the properties of the specific geometric objects involved [9, 2].

*University of Massachusetts Lowell, Lowell MA, USA. Supported by the NSF award CCF-2348067.
†School of Computer Science and Engineering, Ohio University, OH., USA., miirza@ohio.edu
‡Department of Computer Science, Utrecht University, NL.
§Department of Computer Science, Tulane University, LA., USA.

Various approximation algorithms for geometric optimization problems, including set cover and hitting set problems, are proposed, notably in [1, 2], particularly the one that exploits the general method of multiplicative weights update [4].

In the context of covering points with segments under Hausdorff distance constraints, algorithms often involve clustering the points and fitting segments to these clusters. Har-Peled and Mazumdar [5] provided algorithms for clustering in high-dimensional spaces that can be adapted to geometric covering problems.

Problem statement: Given a set S of n points in $\mathbb{R}^d$, and an error $\varepsilon > 0$, find a minimum cardinality set L of line segments such that $\mathsf{H}(S, L) \leq \varepsilon$. where $\mathsf{H}(A, B)$ is the undirected Hausdorff distance between two sets $A, B \subset \mathbb{R}^d$.

Greedy algorithm. We can define a simple greedy algorithm for this problem: incrementally build a solution by choosing a segment in the union of ε-radius disks centered at input points that hits the maximum number of disks that are not yet hit. We can show that such an algorithm cannot do better than a $O(\log n)$-approximation, even if the ply of the disk arrangement (the maximum number of intersections between disks at any given point) is constant. Our lower bound construction is illustrated in Figure 1.

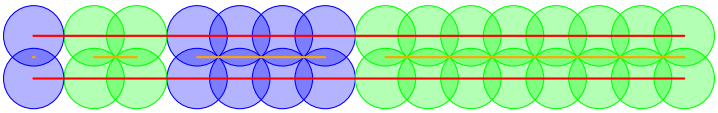


Figure 1: The solution obtained by greedy algorithm is highlighted in orange, and the optimal solution is highlighted in red. This construction uses only 1 orientation, and has ply 2 (or 4, depending on whether we count ply of the open or closed disks).

Observation 1. *Given a points set S with $|S| = n$, let OPT be the cost of the optimal solution for an instance of our problem, and let Gr be the cost of the solution found with the greedy algorithm. Then, in the worst case we have*

$$\mathsf{Gr} = \Omega(\log n)\mathsf{OPT}.$$

2 NP-Hardness Construction using Planar 3-SAT

Reduction. Define the decision version of our problem in the natural way: Given S and ε as previously defined, and $k \in \mathbb{N}$, decide whether a set of line segments L exist such that $\mathsf{H}(S, L) \leq \varepsilon$ and $|L| \leq k$. We reduce from Planar Rectilinear 3-SAT, which is known to be NP-complete [8]. An instance of Planar Rectilinear 3-SAT is given by a boolean formula in 3-CNF with n variables and m clauses, and a planar bipartite incidence graph G of variables and clauses together with an embedding where variables and clauses are mapped to unit-height rectangles, variable rectangles are mapped vertically aligned on the x-axis,

and edges are vertical line segments. We can also assume that all vertices of the drawing lie on integer coordinates.

We modify the planar embedding to obtain an instance of our problem as follows. The construction consists of points whose ε-neighborhoods are interior-disjoint disks. We choose $\varepsilon = 0.8$ so that the majority of the disks lie in the integer positions of the hexagonal grid, as explained in the description of each gadget. The contact graph of these disks will resemble G in the sense that it will contain G as a minor. An example of a reduction is given in Figure 2.

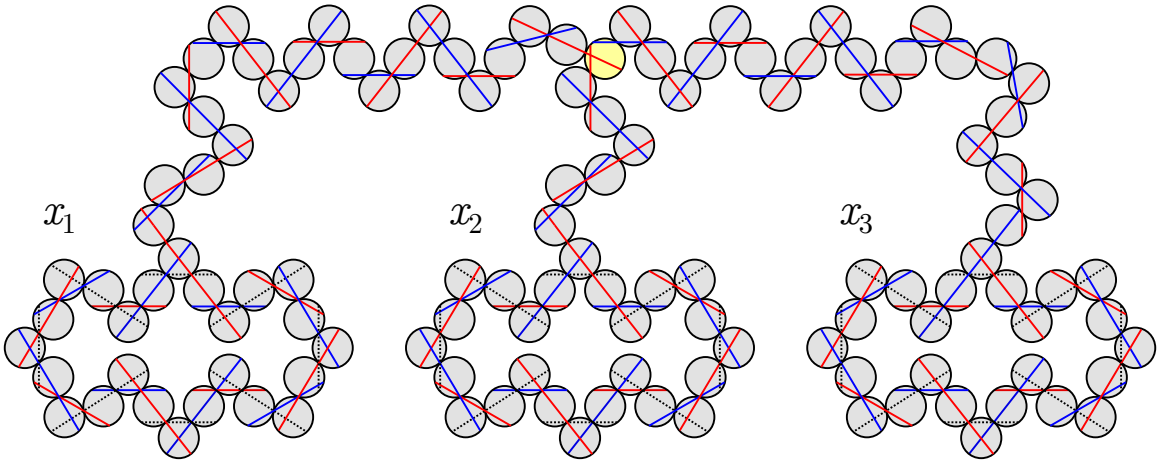
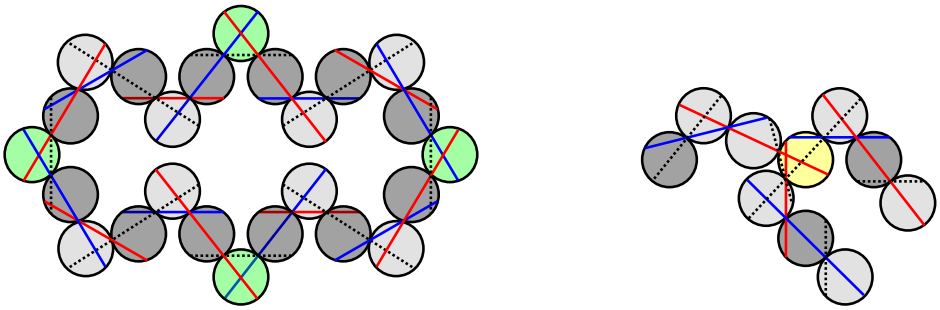


Figure 2: Reduction for the instance $x_1 \vee x_2 \vee \overline{x_3}$.

Replace each rectangle of a variable x_i with a *variable gadget* shown in Figure 3a. The gadget consists of $6(3 + d(x_i))$ disks where $d(x_i)$ denotes the the maximum number of clauses incident to x_i among the upper and lower halfplanes. The disks are arranged so that their contact graph forms a cycle. Note that the number of disks is even and the cycle is bipartite. The coordinate of the centers of the partite class shown in dark gray lie in integer positions. We label the leftmost disk green and do the same for every 6th disk from there. The touching points are so that no three consecutive ones are collinear. In Figure 3a, we also show maximal line segments contained in the disks that contain contain two touching points. By construction, we can split the segments into three independent sets (where two segments are not independent if they intersect) shown in red, blue and dashed black. We label such sets as in the figure, i.e., where the disks containing endpoints of red and blue segments are green.

Place a *clause gadget* in the center of each clause rectangle as shown in Figure 3b. The disks shown in yellow and dark gray have centers in integer positions. The yellow disk is the only touching three other disks. We connect the clause gadgets with paths of disks, called wire gadgets, to a green disk of a corresponding variable gadget as in Figure 2. Note that the disk adjacent to a green disk is positioned so that their tangent point is the endpoint of the red (resp., blue) segment if the variable appears positively (resp., negatively) in the



(a) Variable gadget. (b) Clause gadget.

Figure 3: Gadgets

clause. As in the variable gadget, every other disk in these paths can be positioned in integer position and we can assume that no three consecutive touching points are collinear (except for the disks adjacent to the green disks in variable gadgets). Let d be the number of disks in the instance produced by the reduction. Finally, we set $k = (d - m)/3$, the maximum size of a desired solution of our problem.

Theorem 2. *Given a set of points P in the plane and an $\varepsilon > 0$, deciding whether there is a set of k line segments so that the Hausdorff distance between the sets is at most ε is NP-complete.*

Proof sketch. The problem is in NP since we can compute the Hausdorff distance in polynomial time. Given a solution for the 3SAT instance, choose the red (resp., blue) segments for the variable and wire gadgets corresponding to a variable that is assigned true (resp., false). The yellow disk of a variable gadget is hit by one of the segments coming from a wire gadget. By construction, k segments are used and all disks are hit.

Now consider the other direction. Note that the maximum number of disks that can be hit by a single segment is three, except at the place where variable and wire gadget meets, where a segment can hit four disks. Thus, given a solution to the instance, we can replace any segment to one of the blue, red or dashed black segments shown in the figures. If the solution restricted to a variable gadget with d' disks has $d'/3$ segments, then it must contain segments of a single color class. If the solution contains a black dashed segment and we propagate this signal until the variable gadgets, a green disk will be hit in two variable gadgets. If we propagate the signal while hitting three disks per segment we can show that we will have two disks leftover and therefore need an extra segment. The bound can only then be achieved if we consistently choose a color class for each variable which encodes a satisfying truth assignemnt. □

References

[1] Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. Geometric approximation via coresets. In *Combinatorial and Computational Geometry*, pages 1–30. Cambridge University Press, 2005.

[2] Timothy M. Chan and Qizheng He. Faster Approximation Algorithms for Geometric Set Cover. In Sergio Cabello and Danny Z. Chen, editors, *36th International Symposium on Computational Geometry (SoCG 2020)*, volume 164 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 27:1–27:14, Dagstuhl, Germany, 2020. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[3] Robert J. Fowler, Michael S. Paterson, and Steven L. Tanimoto. Optimal packing and covering in the plane are NP-complete. *Information Processing Letters*, 12(3):133–137, 1981.

[4] Sariel Har-peled. *Geometric Approximation Algorithms*. American Mathematical Society, USA, 2011.

[5] Sariel Har-Peled and S. Mazumdar. Clustering in high dimensions using projected coresets. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing*, pages 641–650. ACM, 2004.

[6] Hugues Hoppe, Tony DeRose, Tom Duchamp, John McDonald, and Werner Stuetzle. Surface reconstruction from unorganized points. In *Proceedings of the 19th annual conference on Computer graphics and interactive techniques*, pages 71–78. ACM, 1992.

[7] Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[8] Donald E Knuth and Arvind Raghunathan. The problem of compatible representatives. *SIAM Journal on Discrete Mathematics*, 5(3):422–427, 1992.

[9] Nabil H. Mustafa and Saurabh Ray. Improved results for geometric set cover. *Discrete & Computational Geometry*, 44(4):883–895, 2010.

[10] Hanan Samet. *Foundations of Multidimensional and Metric Data Structures*. Morgan Kaufmann, 2006.