On k-means for segments and polylines

Sergio Cabello ✉ 🆔
Faculty of Mathematics and Physics, University of Ljubljana, Slovenia
Institute of Mathematics, Physics and Mechanics, Slovenia

Panos Giannopoulos ✉ 🆔
Department of Computer Science, City, University of London, UK

───── **Abstract** ─────

We study the problem of k-means clustering in the space of straight-line segments in $\mathbb{R}^2$ under the Hausdorff distance. For this problem, we give a $(1 + \epsilon)$-approximation algorithm that, for an input of n segments, for any fixed k, and with constant success probability, runs in time $O(n + \varepsilon^{-O(k)} + \varepsilon^{-O(k)} \cdot \log^{O(k)}(\varepsilon^{-1}))$. The algorithm has two main ingredients. Firstly, we express the k-means objective in our metric space as a sum of algebraic functions and use the optimization technique of Vigneron [39] to approximate its minimum. Secondly, we reduce the input size by computing a small size coreset using the sensitivity-based sampling framework by Feldman and Langberg [21, 22]. Our results can be extended to polylines of constant complexity with a running time of $O(n + \varepsilon^{-O(k)})$.

1 Introduction

The *k-means* clustering problem is as follows: Given a point set in a metric space, find a set of points, called *centers*, such that the sum of the squared distances from each input point to its closest center is minimized (over all possible choices of centers). It is a fundamental algorithmic problem with a ubiquitous role in data analysis in numerous application domains. As such, it has been studied extensively in geometric and general metric spaces, under various constraints on the objective and the choice of centers, and with a focus on complexity lower and upper bounds and the quality of the (approximate) solution [1, 2, 3, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 21, 22, 25, 28, 31, 33].

In geometric settings, almost all previous work involves clustering points in some low- or high-dimensional Euclidean space. A notable exception is the relatively recent work on *k-center* [5] and *k-median* [6, 10, 19, 36] clustering for polygonal curves; for *k*-center, one seeks to minimize the maximum distance to the closest center, while for *k*-median, one seeks to minimize just the sum of the distances (instead of the sum of the squares) to the closest centers. We are not aware of any work on k-means clustering for more complex objects than points. In this paper, we consider the k-means problem in the spaces of segments and of polylines of constant complexity in the plane with respect to the Hausdorff distance.

1.1 Formalization of the problem

Let $(\mathcal{S}, d_H)$ be the metric space of closed straight-line segments in $\mathbb{R}^2$, where d_H is the Hausdorff distance. Given a set S of n weighted segments, where each $s \in S$ has an associated positive weight $w_s \in \mathbb{R}_{>0}$, and for any k segments $s_1, \ldots, s_k$ playing the role of

"centers" of the clusters, we define the objective function

$$\text{cost}_S(\{s_1, \ldots, s_k\}) \; := \; \sum_{s \in S} w_s \cdot \min\{d_H^2(s_1, s), \ldots, d_H^2(s_k, s)\}$$

and define the k-means problem as the problem of finding a minimizer, i.e., a set of segments $S^* = \{s_1^*, \ldots, s_k^*\}$ that minimizes the above cost. Note here that quite often we deal with *unweighted* input segments. However, for technical reasons (made clear later in our discussion) in order to incorporate coresets in our algorithm, we have to consider the more general case of weighted segments. Also note that we study the *continuous* version of the problem, where the solution segments can come from anywhere in $(\mathcal{S}, d_H)$. This is harder than the so-called *discrete* version, where the solution segments have to be selected among the input segments.

We also consider the k-means problem for polylines, each with a bounded number of segments, under the Hausdorff distance, where the definition of the problem is analogous.

We remark here on an interesting connection to the older and closely related concept of the *Fréchet mean* [24]. This is a generalization of the classic notion of mean or average to any abstract metric space. For a finite point set P in a metric space $(\mathcal{M}, d)$, a Fréchet mean is any minimizer of the so-called *Fréchet variance* $\text{cost}_P(q) := \sum_{p \in P} d^2(q, p)$, taken over all $q \in \mathcal{M}$. For Euclidean spaces, the Fréchet mean is the usual arithmetic mean. (Other usual means can be recovered as Fréchet means by considering other distances.) The Fréchet mean is a well-studied concept in Statistics and in Riemannian spaces, where sometimes it is known as Karcher mean, see [37] for a general, comprehensive treatment. Computing a Fréchet-mean is precisely the 1-means clustering problem while the k-means is the generalization where the cost of each cluster is given by the functional defining the Fréchet mean.

1.2   Results

Our main result is a $(1 + \varepsilon)$-approximation algorithm for the k-means problem in $(\mathcal{S}, d_H)$. The algorithm runs in $O\left(\left(n + \varepsilon^{-16k+4-\eta} + \varepsilon^{-12k-3}\log^{4k+1}(\varepsilon^{-1})\right)(\log(1/\delta))\right)$ time, for any fixed k, any $\eta > 0$, and with success probability at least $1 - \delta$ (the constant hidden in the O-notation depends on η and k).

There are two main ingredients in our algorithm. For the first one, described in Section 3, we express the k-means objective in the space $(\mathcal{S}, d_H)$ as a sum of algebraic functions of constant description complexity. This algebraic approach allows us to use the optimization technique of Vigneron [39] for approximating the minimum. This is, to the best of our knowledge, the first application of this technique in the context of clustering. While this technique is very expensive when applied directly to the entire set of input segments, we can decrease the running time dramatically by combining it with coresets. This is the second ingredient of our algorithm, described in Section 4, namely, we use the sensitivity framework of Feldman and Langberg [21, 22] to compute a small coreset of the input and then we apply the former algebraic approach to the coreset.

We then extend this result to polylines of description complexity at most $\ell = O(1)$. In this context, each input polyline and each solution polyline has at most ℓ segments, but we may put in the solution polylines that are not part of the input. The running time becomes $O\left(\left(n + \varepsilon^{-O(k\ell)}\right)\log(1/\delta)\right)$.

We start with a side-result, given in Section 2, where we consider the Fréchet mean (or 1-means) problem in a concrete example with two perpendicular segments that intersect at their centers. We show that even in this simple setting the set of Fréchet means is surprisingly complex. The optimum is attained in a 3-dimensional subset of the 4-dimensional parameter

space needed to model the space of candidate segments. This example also prompts to the benefit of looking into an algebraic approach for the general setting.

1.3 Related work

For general metric spaces, k-means (as well as k-median) clustering is APX-hard (when k is part of the input) [15, 26]. Several polynomial-time, constant factor approximation algorithms are known for both the continuous and discrete versions of the problem [1, 8]. For the discrete version, there even exist algorithms that achieve factors arbitrarily close to the lower bound [26] and run in FPT-time with respect to k and the approximation error ε [17].

The Euclidean k-means, where the input is a set of points in $\mathbb{R}^d$, is NP-hard for $d = 2$ [33] and APX-hard when $d = \omega(\log n)$ [2]. The problem admits EPTASs with respect to k and ε [31] and with respect to d and ε [11, 13].

As for the Fréchet mean, it has been considered for persistence diagrams [35, 38], point sets on the unit circle [7], and in the space of graphs [23, 30, 34], to name a few metric spaces far from the Euclidean setting.

1.4 Definitions and notation

For each point $p \in \mathbb{R}^2$, we use $x(p)$ and $y(p)$ for its two coordinates. Thus, $p = (x(p), y(p))$. For any two points $p, q \in \mathbb{R}^2$, we denote by pq the segment with endpoints p and q, and by $|pq|$ the Euclidean distance between them: $|pq|^2 = (x(p) - x(q))^2 + (y(p) - y(q))^2$. For simplicity we assume that all input segments have positive length.

Recall that the Hausdorff distance $d_H(A, B)$ between any two closed subsets $A, B \subset \mathbb{R}^2$ is defined by

$$d_H(A, B) = \max\Big\{\max_{a \in A} \min_{b \in B} |ab|, \ \max_{b \in B} \min_{a \in A} |ab|\Big\}.$$

Define $\delta(a, B) = \min_{q \in B} |ab|$ for the (directed) distance from a point a to a closed set B. It is well known and easy to see that for any two segments $s_1 = a_1 b_1$ and $s_2 = a_2 b_2$ in $\mathcal{S}$

$$d_H(s_1, s_2) = \max\{\delta(a_1, s_2), \delta(b_1, s_2), \delta(a_2, s_1), \delta(b_2, s_1)\}. \tag{1}$$

2 An example of Fréchet mean in $(\mathcal{S}, d_H)$

Let $s_1 = a_1 b_1$, $s_2 = a_2 b_2$ be perpendicular segments, centered at the origin o, with $|s_1| = |s_2| = 2$; see Figure 2. We show that the set of Fréchet means in this case is 3-dimensional. Since a candidate Fréchet mean is a segment, described by four parameters, this example shows that even in relatively simple cases the space of Fréchet means may have a complex structure and the maximum is attained at a large subset that is described algebraically.

Let D be the disk centered at o with radius 1 and let $D_1, \ldots, D_4$ be the disks centered at $a_1/2, a_2/2, b_1/2, b_2/2$ with radius $\frac{1}{2}$. The region $D \setminus (D_1 \cup \cdots \cup D_4)$ has four connected regions. Let A_i be the closure of the connected region in the i-th quadrant; see Figure 2.

▶ **Theorem 1.** *A segment is a Fréchet mean of $\{s_1, s_2\}$ if and only if it goes through the origin and has its endpoints in different regions $A_1, \ldots, A_4$.*

Before proving the theorem, we show the following technical statement.

▶ **Lemma 2.** *Consider any two segments ab and $a'b'$ in the plane. Let $a''b''$ be the segment obtained by translating $a'b'$ such that its center coincides with the center of ab. Then*
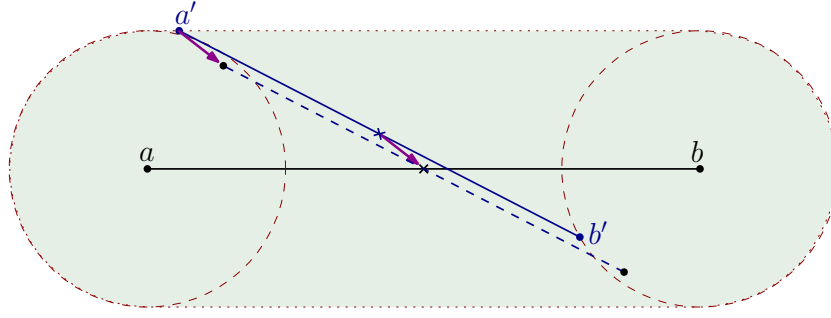
Figure 1 Translating $a'b'$ so that the centers of ab and $a'b'$ coincide. The shaded region contains the points of the plane at distance at most $\delta(a', ab)$ from ab.

$d_H(ab, a''b'') \leq d_H(ab, a'b')$. *Furthermore, if the center of ab does not lie on $a'b'$ and the center of $a'b'$ does not lie on ab, then $d_H(ab, a''b'') < d_H(ab, a'b')$.*

Proof. Without loss of generality we may assume that $\delta(a', ab) \geq \delta(b', ab)$. See Figure 1. Note that

$$\delta(a'', ab) \leq \delta(a', ab),$$

and equality can occur only when the center of $a'b'$ lies on ab. Because of symmetry we have $\delta(a'', ab) = \delta(b'', ab)$. Therefore

$$\max\{\delta(a'', ab), \delta(b'', ab)\} = \delta(a'', ab) \leq \delta(a', ab) = \max\{\delta(a', ab), \delta(b', ab)\},$$

and equality can occur only when the center of $a'b'$ lies on ab. Exchanging the roles of ab and $a'b'$, we also have

$$\max\{\delta(a, a''b''), \delta(b, a''b'')\} \leq \max\{\delta(a, a'b'), \delta(b, a'b')\},$$

and equality can occur only when the center of ab lies on $a'b'$. We then have

$$\begin{aligned} d_H(ab, a'b') &= \max\{\delta(a, a'b'), \delta(b, a'b'), \delta(a', ab), \delta(b', ab)\} \\ &\geq \max\{\delta(a, a''b''), \delta(b, a''b''), \delta(a'', ab), \delta(b'', ab)\} \\ &= d_H(ab, a''b''). \end{aligned}$$

Moreover, when the center of ab does not lie on $a'b'$ and the center of $a'b'$ does not lie on ab, the inequality is strict.   ◀

Proof of Theorem 1. Note that $d_H(s_1, s_2) = 1$. Therefore, for any solution (Fréchet mean) segment s we have $d_H^2(s, s_1) + d_H^2(s, s_2) \leq d_H^2(s_1, s_1) + d_H^2(s_1, s_2) = 0 + d_H^2(s_1, s_2) = 1$.

Observe that any solution segment s must pass through the origin. Indeed, if such a solution s does not pass through the origin, then the center of s does not lie on s_1 or it does not lie on s_2. It follows from Lemma 2 that, by translating s such that its center coincides with the origin, both $d_H(s, s_1)$ and $d_H(s, s_2)$ do not increase, and at least one of them strictly decreases. Therefore, $d_H^2(s, s_1) + d_H^2(s, s_2)$ strictly decreases by translating s such that its center coincides with the origin.

Next, consider a solution segment $s = ab$ that passes through the origin o with its endpoints a, b in the first and third quadrant, respectively. See Figure 2. Let a_c and a' be the points of intersection of the supporting line ℓ of s with the boundary of D and the
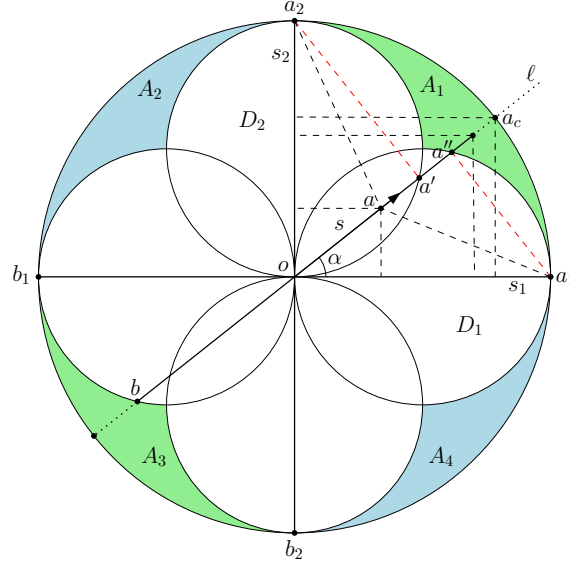
Figure 2 For the set of segments $S = \{s_1, s_2\}$, there is an infinite family of Fréchet means.

boundary of D_2 respectively. Note that $\angle oa'a_2$ is right. As a moves from o to a', we have $\delta(a_2, s) = |a_2 a| > \delta(a, s_2)$ and the unique minimum of $|a_2 a|$ is attained at $a = a'$. When a moves from a' to a_c, $\delta(a_2, s) = |a_2 a'|$ is constant. At $a = a_c$ we have that $\delta(a_2, s) = \delta(a_c, s_2)$, and beyond that $\delta(a_2, s) < \delta(a, s_2)$. Similarly, the minimum value of $\delta(a_1, s)$ is $|a_1 a''|$, where a'' is the projection of a_1 onto ℓ (the intersection of ℓ with the boundary of D_1). We also have that $\delta(a_1, s) = \delta(a_c, s_1)$ when a lies between a'' and a_c, and beyond a_c, $\delta(a_1, s) < \delta(a, s_1)$.

Assume that $a'' \in a'a_c$, as in Figure 2; the other case is symmetric. Using $\alpha := \angle aoa_1$, for every position of a in the segment $a''a_c$ we have

$$\delta^2(a, s_1) + \delta^2(a, s_2) \ \leq \ \delta^2(a_1, s) + \delta^2(a_2, s) \ = \ |a_1 a''|^2 + |a_2 a'|^2 \ = \ \sin^2 \alpha + \sin^2(\pi/2 - \alpha) \ = \ 1.$$

For every position of a in the segment oa'' we have $\delta^2(a_1, s) + \delta^2(a_2, s) \ > \ |a_1 a''|^2 + |a_2 a'|^2 \ = \ 1$, while for every position of a past a_c we have that $\delta^2(a, s_1) + \delta^2(a, s_2) > \delta^2(a_1, s) + \delta^2(a_2, s) = 1$.

For the endpoint b of s in the third quadrant, the situation is symmetric. Using that the Hausdorff distance is attained at some endpoint, we get that each segment with an endpoint in A_1 and an endpoint in A_3 gives the minimum possible value of $d_H^2(s, s_1) + d_H^2(s, s_2) \geq 1$, and therefore is a Fréchet mean. The situation is symmetric for the other quadrants. ◀

3 An algebraic approach to k-means in $(\mathcal{S}, d_H)$

We use the following adaptation of the definition of a *nice* family of functions by Vigneron [39, Section 2.1]. Let $\mathcal{F} = \{f_i : \mathbb{R}^d \to \mathbb{R} \mid i \in I\}$ be a finite family of functions, where I is some index set. We say that $\mathcal{F}$ is *nice* if there exists a constant $\lambda > d > 0$ such that:

- each $f_i \in \mathcal{F}$ is nonnegative and bounded;
- for each $f_i \in \mathcal{F}$, there exists a semialgebraic set $\text{supp}(f_i) \subseteq \mathbb{R}^d$ and an algebraic function g_i of degree at most λ with $f_i(x) = g_i(x)$ for $x \in \text{supp}(f_i)$ and $f_i(x) = 0$ for $x \notin \text{supp}(f_i)$;
- for each $f_i \in \mathcal{F}$, the semialgebraic set $\text{supp}(f_i) \subseteq \mathbb{R}^d$ is a boolean combination of at most λ subsets of $\mathbb{R}^d$, each of them defined by an polynomial inequality of degree at most λ;
- for each $f_i \in \mathcal{F}$, the restriction of f_i to $\text{supp}(f_i)$ is continuous.

Note that the definition allows that the sets $\text{supp}(f_i)$ are open, closed or mixed. It also allows that f_i is discontinuous in $\mathbb{R}^d \setminus \text{supp}(f_i)$, which may include the boundary of $\text{supp}(f_i)$ in some cases.

Our use of this concept will be through the following result for computing an approximation to the minimum of the function $\sum_i f_i$.

▶ **Theorem 3** (Adaptation of Theorem 3.4 in Vigneron [39]). *Assume that $\varepsilon \in (0, 1)$. Let $\mathcal{F} = \{f_i : \mathbb{R}^d \to \mathbb{R} \mid i \in I\}$ be a nice family of m functions. Define $g = \sum_{i \in I} f_i$ and assume that $\min_{x \in \mathbb{R}^d} g(x)$ exists. Then we can compute a point $x'_\varepsilon \in \mathbb{R}^d$ such that $g(x'_\varepsilon) \leq (1 + \varepsilon) \min_{x \in \mathbb{R}^d} g(x)$ in time $O(m^{2d-2+\eta} + (m/\varepsilon)^{d+1} \log^{d+1}(m/\varepsilon))$ for any $\eta > 0$.[1] The constant hidden in the O-notation depends on η and on d.*

Let us first consider the simpler case of two segments and how their Hausdorff distance is defined. We parameterize a segment ab as the point $(x(a), y(a), x(b), y(b))$ in $\mathbb{R}^4$. Note that in this parameterization we have the artifact that the segments ab and ba give different points in $\mathbb{R}^4$.

Let ℓ_{ab} be the line supporting a segment ab. For a point p, the distance $\delta(p, ab)$ is given by one of the three terms $|pa|$, $|pb|$, or $\delta(p, \ell_{ab})$. For a point $q \in \mathbb{R}^2$ and a segment ab, let $\ell_\perp(q, ab)$ be the line perpendicular to ℓ_{ab} through q. The lines $\ell_\perp(a, ab)$ and $\ell_\perp(b, ab)$ partition the plane into three 2-dimensional faces (Figure 3) with closures

$$\sigma(ab) = \text{the closed slab between } \ell_\perp(a, ab) \text{ and } \ell_\perp(b, ab),$$
$$\tau(a, ab) = \text{the closed halfspace defined by } \ell_\perp(a, ab) \text{ that does not contain } b,$$
$$\tau(b, ab) = \text{the closed halfspace defined by } \ell_\perp(b, ab) \text{ that does not contain } a.$$

We then have

$$\delta(p, ab) = \begin{cases} |pa| & \text{if } p \in \tau(a, ab), \\ |pb| & \text{if } p \in \tau(b, ab), \\ \delta(p, \ell_{ab}) & \text{if } p \in \sigma(ab). \end{cases}$$

From Equation (1), we conclude that, for any two segments ab and $a'b'$, the distance $d_H(ab, a'b')$ is given by one of the functions in the family

$$\mathcal{F}(ab, a'b') := \big\{|aa'|, \ |ab'|, \ |ba'|, \ |bb'|, \ \delta(a, \ell_{a'b'}), \ \delta(b, \ell_{a'b'}), \ \delta(a', \ell_{ab}), \ \delta(b', \ell_{ab})\big\}.$$

We next argue that all the expressions involved are algebraic. A point p lies on the line $\ell_\perp(a, ab)$ if and only the scalar product of the vectors $\vec{ap}$ and $\vec{ab}$ is zero. This is equivalent to $\big(x(p), y(p), x(a), y(a), x(b), y(b)\big)$ being a zero of the algebraic (actually polynomial) function

$$\psi(x, y, x_a, y_a, x_b, y_b) := (x - x_a)(x_b - x_a) + (y - y_a)(y_b - y_a).$$

The sign of this expression also tells us which side of $\ell_\perp(a, ab)$ the point p lies on. Note that this function is linear in x and y, while it is quadratic in x_a and y_a. Symmetrically, the sign of $\psi\big(x(p), y(p), x(b), y(b), x(a), y(a)\big)$ tells us which side of $\ell_\perp(b, ab)$ point p lies on.

In the following, we will treat the segment $a'b'$ as variable, identified with $\mathbb{R}^4$, while the segment ab will be fixed. We will show that the space $\mathbb{R}^4$ can be decomposed into cells such

---

[1] As noted in Vigneron [39], one needs to assume either the Real-RAM model of computation (which is standard in computational geometry) or a model where we can choose the precision of the intermediate computations, and then the computational complexity of the algorithm depends on the desired precision.
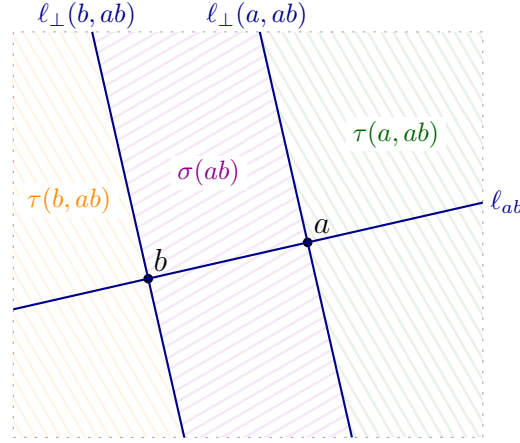
Figure 3 The regions $\sigma(ab)$, $\tau(a, ab)$ and $\tau(b, ab)$.

that, within a cell, the distance $d_H(ab, a'b')$ is defined always by the same function from $\mathcal{F}(ab, a'b')$. Such a decomposition is given by the eight algebraic hypersurfaces describing the conditions

$$a' \in \ell_\perp(a, ab), \ b' \in \ell_\perp(a, ab), \ a' \in \ell_\perp(b, ab), \ b' \in \ell_\perp(b, ab),$$
$$a \in \ell_\perp(a', a'b'), \ b \in \ell_\perp(a', a'b'), \ a \in \ell_\perp(b', a'b'), \ b \in \ell_\perp(b', a'b'),$$

together with a set of hypersurfaces, "bisectors", each defined by the set of points where two appropriate functions from $\mathcal{F}(ab, a'b')$ meet; this will become clear shortly. Finally, we note that each function in $\mathcal{F}(ab, a'b')$ is algebraic of constant degree; for example, elementary algebra shows that

$$\delta^2(a', \ell_{ab}) \ = \ \frac{\Big(\big(x(b) - x(a)\big)\big(y(a) - y(a')\big) - \big(x(a) - x(a')\big)\big(y(b) - y(a)\big) \Big)^2}{\big(x(a) - x(b)\big)^2 + \big(y(a) - y(b)\big)^2}.$$

We parameterize the space of (sequences of) k segments $a_1b_1, \ldots, a_kb_k$ (the k candidate cluster centers) by the point

$$\big(x(a_1), y(a_1), x(b_1), y(b_1), \ldots, x(a_k), y(a_k), x(b_k), y(b_k)\big) \in \mathbb{R}^{4k}.$$

Similarly, each $z \in \mathbb{R}^{4k}$ defines a k-tuple of segments with $s_1(z) = a_1(z)b_1(z), \ldots, s_k(z) = a_k(z)b_k(z)$ by taking the inverse of the parameterization.

▶ **Theorem 4.** *Let k be a fixed, positive integer and let s be a segment in the plane. In $O(1)$ time we can construct a nice family $\mathcal{F}_s = \{ f : \mathbb{R}^{4k} \to \mathbb{R} \}$ of $O(1)$ functions such that*

$$\forall z \in \mathbb{R}^{4k} : \quad \sum_{f \in \mathcal{F}_s} f(z) = \min_{i \in [k]} d_H^2(s, s_i(z)).$$

Proof. Let $s = ab$ be the fixed segment. For each index $i \in [k]$, we consider the set $\Sigma(i)$ of 8 hypersurfaces in $\mathbb{R}^{4k}$, each of them given by one of the following conditions

$$a_i \in \ell_\perp(a, ab), \ b_i \in \ell_\perp(a, ab), \ a_i \in \ell_\perp(b, ab), \ b_i \in \ell_\perp(b, ab),$$
$$a \in \ell_\perp(a_i, a_ib_i), \ b \in \ell_\perp(a_i, a_ib_i), \ a \in \ell_\perp(b_i, a_ib_i), \ b \in \ell_\perp(b_i, a_ib_i).$$

Note that here $x(a)$, $y(a)$, $x(b)$ and $y(b)$ are input data while $x(a_i)$, $y(a_i)$, $x(b_i)$ and $y(b_i)$ are variables defining coordinates in the parameter space $\mathbb{R}^{4k}$.

Set $\Sigma := \cup_{i \in [k]} \Sigma(i)$ and let $\mathcal{A}_\Sigma$ be the arrangement in $\mathbb{R}^{4k}$ defined by Σ. From the foregoing discussion, we have the following property: for each cell c of $\mathcal{A}_\Sigma$ and each index $i \in [k]$, the set of functions $\mathcal{F}(s, a_i b_i)$ stays the same and each of the distances $\delta(a, a_i b_i)$, $\delta(b, a_i b_i)$, $\delta(a_i, ab)$, and $\delta(b_i, ab)$, is given by the same function from $\mathcal{F}(s, a_i b_i)$. Thus, for all $z \in c$, the distance $d_H(s, s_i(z))$ is described by the maximum among the same four functions from $\mathcal{F}(s, a_i b_i)$.

In order to make clear that only the coordinates of a_i and b_i are relevant in the functions in $\mathcal{F}(s, a_i b_i)$, we change the notation to $\mathcal{G}_i$ and take each function g of $\mathcal{G}_i$ to map from $\mathbb{R}^{4k}$ to $\mathbb{R}$. Formally, for each function $f \in \mathcal{F}(s, a_i b_i)$ we put into $\mathcal{G}_i$ the function $g(z) := f(ab, s_i(z))$.

We next define a set Λ of algebraic hypersurfaces in $\mathbb{R}^{4k}$ playing the role of "bisectors". For each $i, j \in [k]$ with $i \leq j$, we define $\Lambda(i, j)$ as the hypersurfaces given by equating each function of $\mathcal{G}_i$ to each function of $\mathcal{G}_j$. Note that each hypersurface is defined by a polynomial equality of degree at most 6. Since $\mathcal{G}_i$ has 8 functions for each $i \in [k]$, the set $\Lambda(i, j)$ has at most $8^2 = 64$ hypersurfaces (it is 32 for $\Lambda(i, i)$).

Set $\Lambda := \cup_{i \in [k]} \cup_{j \in [k], i \leq j} \Lambda(i, j)$ and let $\mathcal{A}_\Lambda$ be the arrangement in $\mathbb{R}^{4k}$ induced by Λ. For each cell $c \in \mathcal{A}_\Lambda$ the sign of each function $g(z) - g'(z)$ remains constant for $g \in \mathcal{G}_i$, $g' \in \mathcal{G}_j$ and $z \in c$.

Finally, let $\mathcal{A}$ be the arrangement in $\mathbb{R}^{4k}$ induced by the hypersurfaces in $\Sigma \cup \Lambda$. Note that this is a refinement of $\mathcal{A}_\Sigma$ and $\mathcal{A}_\Lambda$, meaning that each cell of $\mathcal{A}$ is contained in a cell of $\mathcal{A}_\Sigma$ and a cell of $\mathcal{A}_\Lambda$.

Consider a cell $c \in \mathcal{A}$. Since c is contained in a cell of $\mathcal{A}_\Sigma$, for each $i \in [k]$, each function in the set $\Delta_i(c) = \{\delta(a, s_i(z)), \delta(b, s_i(z)), \delta(a_i(z), ab), \text{ and } \delta(b_i(z), ab)\}$ is given by the same function of $\mathcal{G}_i$ for all $z \in c$. Moreover, since c is contained in a cell of $\mathcal{A}_\Lambda$, for every two distinct functions $\delta, \delta' \in \Delta_i(c)$ the sign of $\delta - \delta'$ is constant for all $z \in c$. From these two facts we conclude that, for each $i \in [k]$, there is some function $g_{c,i}(z) \in \mathcal{G}_i$ such that $d_H(s, s_i(z)) = g_{c,i}(z)$ for all $z \in c$. This function can be easily determined in $O(1)$ time per cell by evaluating each function in $\Delta_i(c)$ at some arbitrary point in c.

Similarly, since c is contained in a cell of $\mathcal{A}_\Lambda$, we have that for each distinct $i, j \in [k]$ the sign of

$$d_H(s, s_i(z)) - d_H(s, s_j(z)) \; = \; g_{c,i}(z) - g_{c,j}(z)$$

is constant for all $z \in c$. This implies that, for each cell $c \in \mathcal{A}$, there exists some index $\iota(c) \in [k]$ with the following property:

$$\forall j \in [k], \; z \in c : \quad d_H(s, s_{\iota(c)}(z)) \leq d_H(s, s_j(z)).$$

In other words, the segment $s_{\iota(c)}(z)$ is a closest one to s among $s_1(z), \ldots, s_k(z)$ and moreover the distance $d_H(s, s_{\iota(c)}(z))$ is given by a single function $g_{c,\iota(c)}$ from $\mathcal{G}_{\iota(c)}$. Thus, for each $z \in c$ it holds $\min_{i \in [k]} d_H(s, s_i(z)) = g_{c,\iota(c)}(z)$. As before, the function $g_{c,\iota(c)}(z)$ can be determined in $O(k)$ time per cell by evaluating each $d_H(s, s_i(z))$ at some arbitrary point in c.

For any set A, let 1_A be the function with $1_A(x) = 1$ if $x \in A$ and $1_A(x) = 0$ if $x \notin A$. For each cell $c \in \mathcal{A}$, define the function $h_c : \mathbb{R}^{4k} \to \mathbb{R}$ by $h_c(z) = 1_c(z) \cdot g_{c,\iota(c)}^2(z)$. Finally, set $\mathcal{F}_s := \{h_c \mid c \in \mathcal{A}\}$. We can then express the function

$$z \in \mathbb{R}^{4k} \mapsto \min_{i \in [k]} d_H^2(s, s_i(z))$$

as

$$\min_{i \in [k]} d_H^2(s, s_i(z)) \; = \; \sum_{c \in \mathcal{A}} 1_c(z) \, g_{c,\iota(c)}^2(z) \; = \; \sum_{c \in \mathcal{A}} h_c(z) \; = \; \sum_{h \in \mathcal{F}_s} h(z).$$

Since $\Sigma \cup \Lambda$ has $O(k^2) = O(1)$ hypersurfaces, the arrangement $\mathcal{A}$ has $O(O(k^2)^{4k}) = O(1)$ cells, each of them described by $O(k^2) = O(1)$ algebraic inequalities of constant description complexity and the family of functions $\mathcal{F}_s$ has the desired properties, where the constant λ used to define the niceness is $O(k^{8k})$. Constructing $\mathcal{A}$ (i.e., with algebraic descriptions for each cell) takes $O(O(k^2)^{4k+1}6^{O((4k)^4)}) = O(1)$ [4, Chapter 16]. The family $\mathcal{F}_s$ can be constructed in this time as well. ◀

We can now apply Theorem 3 combining all the functions $\mathcal{F}_s$ for $s \in S$ and compute a set of k segments whose cost approximates that of an optimal set of segments.

▶ **Theorem 5.** *Let k a fixed, positive integer and let $\varepsilon \in (0,1)$. Let S be a family of n segments in the plane with positive weights. We can compute k segments $s_{1,\varepsilon}, \ldots, s_{k,\varepsilon}$ in $\mathbb{R}^2$ such that*

$$\mathrm{cost}_S(\{s_{1,\varepsilon}, \ldots, s_{k,\varepsilon}\}) \leq (1+\varepsilon) \min\Big\{\mathrm{cost}_S(\{s_1, \ldots, s_k\}) \mid s_1, \ldots, s_k \ segments\Big\}.$$

in time $O(n^{8k-2+\eta} + (n/\varepsilon)^{4k+1} \log^{4k+1}(n/\varepsilon))$, for any $\eta > 0$. The constant hidden in the O-notation depends on η and on k.

Proof. For each segment $s \in S$ we compute the family $\mathcal{F}_s$ of Theorem 4. This takes $O(n) \cdot O(1) = O(n)$ time in total. To account for the weight $w_s > 0$ of the segment s, we replace in $\mathcal{F}_s$ each function $f \in \mathcal{F}_s$ with $w_s \cdot f$. Define $\mathcal{F} := \cup_{s \in S} \mathcal{F}_s$ and the function $g := \sum_{f \in \mathcal{F}} f$. Note that $\mathcal{F}$ is a family of $O(1) \cdot O(n) = O(n)$ nice functions and

$$\forall z \in \mathbb{R}^{4k}: \ \ g(z) = \sum_{s \in S} \sum_{f \in \mathcal{F}_s} f(z) = \sum_{s \in S} w_s \cdot \min_{i \in [k]} d_H(s, s_i(z))^2 = \mathrm{cost}_S(\{s_1(z), \ldots, s_k(z)\}).$$

We can then use Theorem 3 to find in time $O(|\mathcal{F}|^{2 \cdot 4k-2+\eta} + (|\mathcal{F}|/\varepsilon)^{4k+1} \log^{4k+1}(|\mathcal{F}|/\varepsilon))$, for any $\eta > 0$, a point $z'_\varepsilon \in \mathbb{R}^{4k}$ such that

$$g(z'_\varepsilon) \leq (1+\varepsilon) \min_{z \in \mathbb{R}^{4k}} \mathrm{cost}_S(\{s_1(z), \ldots, s_k(z)\}).$$

The point $z'_\varepsilon \in \mathbb{R}^{4k}$ defines the segments $s_{1,\varepsilon} := s_1(z'_\varepsilon), \ldots, s_{k,\varepsilon} := s_k(z'_\varepsilon)$. Since $s_1(z), \ldots, s_k(z)$ goes over all k tuples of segments when z iterates over all $\mathbb{R}^{4k}$, we have

$$\min_{z \in \mathbb{R}^{4k}} \mathrm{cost}_S(\{s_1(z), \ldots, s_k(z)\}) = \min_{s_1, \ldots, s_k} \mathrm{cost}_S(\{s_1, \ldots, s_k\}).$$

We conclude that

$$\mathrm{cost}_S(\{s_{1,\varepsilon}, \ldots, s_{k,\varepsilon}\} = g(z'_\varepsilon) \leq (1+\varepsilon) \min_{s_1, \ldots, s_k} \mathrm{cost}_S(\{s_1, \ldots, s_k\}). \qquad ◀$$

4   A coreset for k-means in $(\mathcal{S}, d_H)$

We use the sensitivity framework of Feldman and Langberg [21, 22]. Let F be a finite set of functions, each of them mapping from $\mathbb{R}^d$ to $\mathbb{R}_{\geq 0}$. The *sensitivity* of $f \in F$ with respect to F is

$$\sigma(f, F) := \sup_{z \in \mathbb{R}^d} \frac{f(z)}{\sum_{g \in F} g(z)}.$$

We also consider the following range space

$$\mathrm{range}_{\geq}(F) := \Big(F, \big\{\{f \in F \mid f(z) \geq r\} \mid z \in \mathbb{R}^d, \ r \in [0, \infty)\big\}\Big).$$

We will use the following theorem from [22], which we state here adapted to our needs.

▶ **Theorem 6** (Adaptation of Theorem 31 in Feldman et al. [22])**.** *Let F be a set of n functions from $\mathbb{R}^d$ to $[0, \infty)$ with the following properties:*

- *For each choice of weights $w_f > 0$ for $f \in F$, the range space $\mathrm{range}_{\geq}(\{w_f \cdot f \mid f \in F\})$ has bounded VC-dimension.*
- *For each $f \in F$ we are given a value $\tilde{\sigma}(f)$ such that*

$$\tilde{\sigma}(f) \geq \frac{1}{|F|} \quad and \quad \tilde{\sigma}(f) \geq \sigma(f, F).$$

Set $\tilde{\Sigma}(F) := \sum_{f \in F} \tilde{\sigma}(f)$. Let δ, ε be real values in $(0, 1/2)$. In time $O(|F|)$ we can compute a subset $C \subseteq F$ of

$$O\left(\frac{\tilde{\Sigma}(F)}{\varepsilon^2} \left(\log \tilde{\Sigma}(F) + \log \frac{1}{\delta} \right) \right)$$

weighted functions and weights $u_f > 0$ for each $f \in C$ such that, with probability at least $1 - \delta$:

$$\forall z \in \mathbb{R}^d : \quad \left| \sum_{f \in F} f(z) - \sum_{f \in C} u_f \cdot f(z) \right| \leq \varepsilon \sum_{f \in F} f(z).$$

For each input segment $s \in S$, we define the function $f_s : \mathbb{R}^{4k} \to \mathbb{R}_{\geq 0}$ with

$$f_s(z) := \min\{d_H^2(s, s_1(z)), \ldots, d_H^2(s, s_k(z))\} = \left(\min\{d_H(s, s_1(z)), \ldots, d_H(s, s_k(z))\} \right)^2.$$

Here, the segments $s_1(z), \ldots, s_k(z)$ are the same that were used in the parameterization before Theorem 4. Set $F = \{f_s \mid s \in S\}$. In order to use the above theorem, we need appropriate sensitivity upper bounds $\tilde{\sigma}(f_s)$ for each $f_s \in F$ and a bound on the total sensitivity $\tilde{\Sigma}(F)$. Let $\mathrm{opt}_k(S)$ be the cost of an optimal set of segments for k-means, i.e., $\mathrm{opt}_k(S) = \min_{s_1, \ldots, s_k} \mathrm{cost}_S(\{s_1, \ldots, s_k\})$.

▶ **Lemma 7.** *Let $s_1' \ldots, s_{k'}'$ be a bicriteria (α, β)-approximation for k-means, that is, $k' \leq \beta k$ and $\mathrm{cost}_S(\{s_1', \ldots, s_{k'}'\}) \leq \alpha \cdot \mathrm{opt}_k(S)$, where $\alpha, \beta \geq 1$. For each $i \in [k']$, let S_i' be the segments of S closer to s_i' than to any other segment s_j', $j \in [k'] \setminus \{i\}$; ties are solved arbitrarily so that $S_1', \ldots, S_{k'}'$ is a partition of S. For each segment $s \in S$, let $\iota(s) \in [k']$ be such that $s \in S_{\iota(s)}'$. Define for each $s \in S$ the value*

$$\tilde{\sigma}(f_s) := \frac{32\alpha}{|S_{\iota(s)}'|} + \frac{16\alpha \cdot d_H^2(s, s_{\iota(s)}')}{\sum\limits_{s' \in S_{\iota(s)}'} d_H^2(s', s_{\iota(s)}')} = \frac{32\alpha}{|S_{\iota(s)}'|} + \frac{16\alpha \cdot d_H^2(s, s_{\iota(s)}')}{\mathrm{cost}_{S_{\iota(s)}'}(s_{\iota(s)}')}.$$

Then $\tilde{\sigma}(f_s) \geq \sigma(f_s, F)$ and $\tilde{\sigma}(f_s) \geq \frac{1}{|F|}$.

Proof. Consider a fixed $s \in S$ and set $i = \iota(s)$; thus $s \in S_i'$. Define R_i to be the average cost of the elements of S_i' for the center s_i':

$$R_i = \frac{\mathrm{cost}_{S_i'}(s_i')}{|S_i'|} = \frac{\sum\limits_{s' \in S_i'} d_H^2(s', s_i')}{|S_i'|}.$$

With this notation we have

$$\tilde{\sigma}(f_s) = \frac{32\alpha}{|S_i'|} + \frac{16\alpha \cdot d_H^2(s, s_i')}{\mathrm{cost}_{S_i'}(s_i')} = \frac{32\alpha}{|S_i'|} + \frac{16\alpha \cdot d_H^2(s, s_i')}{|S_i'| \cdot R_i}.$$

Since $|S_i'| \leq |S| = |F|$ and $\alpha \geq 1$, we have $\tilde{\sigma}(f_s) \geq \frac{1}{|S_i'|} \geq \frac{1}{|F|}$.

For the other inequality, recall that

$$\sigma(f_s, F) = \sup_{z \in \mathbb{R}^{4k}} \frac{f_s(z)}{\sum_{f_{s'} \in F} f_{s'}(z)} = \sup_{z \in \mathbb{R}^{4k}} \frac{\min\{d_H^2(s, s_1(z)), \dots, d_H^2(s, s_k(z))\}}{\sum_{s' \in S} \min\{d_H^2(s', s_1(z)), \dots, d_H^2(s', s_k(z))\}}.$$

Consider any $z \in \mathbb{R}^{4k}$. We distinguish three cases.

First, if z is such that for some $j \in [k]$ we have $d_H^2(s, s_j(z)) \leq 16 \cdot d_H^2(s, s_i')$, then we use that $\text{cost}_S(\{s_1', \dots, s_{k'}'\}) \leq \alpha \cdot \text{opt}_k(S)$ to obtain

$$\frac{f_s(z)}{\sum_{s' \in S} f_{s'}(z)} \leq \frac{d_H^2(s, s_j(z))}{\text{opt}_k(S)} \leq \frac{16\alpha \cdot d_H^2(s, s_i')}{\text{cost}_S(\{s_1', \dots, s_{k'}'\})} \leq \frac{16\alpha \cdot d_H^2(s, s_i')}{\text{cost}_{S_i'}(s_i')} \leq \tilde{\sigma}(f_s).$$

Second, if z is such that for some $j \in [k]$ we have $d_H^2(s, s_j(z)) \leq 16 \cdot R_i$, then we use again that $\text{cost}_S(\{s_1', \dots, s_{k'}'\}) \leq \alpha \cdot \text{opt}_k(S)$ to obtain

$$\frac{f_s(z)}{\sum_{s' \in S} f_{s'}(z)} \leq \frac{d_H^2(s, s_j(z))}{\text{opt}_k(S)} \leq \frac{16\alpha \cdot R_i}{\text{cost}_S(\{s_1', \dots, s_{k'}'\})} \leq \frac{16\alpha \cdot R_i}{\text{cost}_{S_i'}(s_i')} = \frac{16\alpha \cdot R_i}{|S_i'| \cdot R_i} = \frac{16\alpha}{|S_i'|}$$

$$\leq \tilde{\sigma}(f_s).$$

It remains to handle the case when, for the $z \in \mathbb{R}^{4k}$ under consideration, we have $d_H^2(s, s_j(z)) > 16 \cdot d_H^2(s, s_i')$ and $d_H^2(s, s_j(z)) > 16 \cdot R_i$ for all $j \in [k]$. Let s_j'' be the segment defined by z that is closest to s, which means that $f_s(z) = d_H^2(s, s_j'')$. We then have $d_H(s, s_j'') > 4 \cdot d_H(s, s_i')$ and $d_H(s, s_j'') > 4 \cdot \sqrt{R_i}$. Because of the triangular inequality for $d_H(\cdot, \cdot)$ and because s_j'' is closest to s, for each segment $s' \in S$ and each $\ell \in [k]$ we have

$$d_H(s, s_j'') \leq d_H(s, s_\ell(z)) \leq d_H(s, s_i') + d_H(s_i', s') + d_H(s', s_\ell(z)).$$

This means that,

$$\forall s' \in S: \quad \min\{d_H(s', s_1(z)), \dots, d_H(s', s_k(z))\} \geq d_H(s, s_j'') - d_H(s, s_i') - d_H(s_i', s').$$

Using that $d_H(s, s_i') < d_H(s, s_j'')/4$ we get:

$$\forall s' \in S: \quad \min\{d_H(s', s_1(z)), \dots, d_H(s', s_k(z))\} \geq \tfrac{3}{4} d_H(s, s_j'') - d_H(s_i', s').$$

At least half of the segments $s' \in S_i'$ must have $d_H^2(s', s_i') \leq 2R_i$; otherwise the average cost of the elements of S_i' would be larger than R_i. Let S_i'' be those segments s' of S_i' with $d_H^2(s', s_i') \leq 2R_i$. We then have

$$\forall s' \in S_i'': \quad \min\{d_H(s', s_1(z)), \dots, d_H(s', s_k(z))\} \geq \tfrac{3}{4} d_H(s, s_j'') - d_H(s_i', s')$$
$$\geq \tfrac{3}{4} d_H(s, s_j'') - \sqrt{2R_i}$$
$$\geq \tfrac{3}{4} d_H(s, s_j'') - \tfrac{1}{2\sqrt{2}} d_H(s, s_j'')$$
$$\geq \tfrac{1}{4} d_H(s, s_j'').$$

and therefore

$$\forall s' \in S_i'': \quad f_{s'}(z) = \left(\min\{d_H(s', s_1(z)), \dots, d_H(s', s_k(z))\}\right)^2 \geq \tfrac{1}{16} d_H^2(s, s_j'').$$

It follows that

$$\frac{f_s(z)}{\displaystyle\sum_{s'\in S} f_{s'}(z)} \;\leq\; \frac{d_H^2(s, s_j'')}{\displaystyle\sum_{s'\in S_i''} f_{s'}(z)} \;\leq\; \frac{d_H^2(s, s_j'')}{|S_i''| \cdot \frac{1}{16} d_H^2(s, s_j'')} \;\leq\; \frac{2\cdot 16}{|S_i'|} \;\leq\; \tilde{\sigma}(f_s).$$

We have shown that

$$\forall z \in \mathbb{R}^{4k} : \quad \frac{f_s(z)}{\displaystyle\sum_{s'\in S} f_{s'}(z)} \;\leq\; \tilde{\sigma}(f_s).$$

It then follows that

$$\sigma(f_s, F) \;=\; \sup_{z\in\mathbb{R}^{4k}} \frac{f_s(z)}{\displaystyle\sum_{f_{s'}\in F} f_{s'}(z)} \;\leq\; \tilde{\sigma}(f_s). \hspace{2cm} \blacktriangleleft$$

Finally, note that for the sensitivities $\tilde{\sigma}(f_s)$ defined in Lemma 7, we have the total sensitivity

$$
\begin{aligned}
\tilde{\Sigma}(F) \;&=\; \sum_{s\in S} \tilde{\sigma}(f_s) \;=\; \sum_{s\in S}\left(\frac{32\alpha}{|S_{\iota(s)}'|} + \frac{16\alpha\cdot d_H^2(s, s_{\iota(s)}')}{\mathrm{cost}_{S_{\iota(s)}'}(s_{\iota(s)}')} \right) \\
&=\; \sum_{i\in[k']}\left(\sum_{s\in S_i'} \frac{32\alpha}{|S_i'|} + \sum_{s\in S_i'} \frac{16\alpha\cdot d_H^2(s, s_i')}{\mathrm{cost}_{S_i'}(s_i')} \right) \;=\; \sum_{i\in[k']} (32\alpha + 16\alpha) \\
&=\; O(\beta\alpha k).
\end{aligned}
$$

Next, we bound the VC-dimension of the range space associated with the input segments.

▶ **Lemma 8.** *Assume that we have a weight $w_s > 0$ for each $s \in S$ and consider the set of functions $F_w = \{w_s \cdot f_s \mid s \in S\}$. The range space $\mathrm{range}_\geq(F_w)$ has VC-dimension $O(1)$.*

Proof. First note that the range space $\mathrm{range}_\geq(F_w)$ is equivalent to the range space (S, R), where the ranges are

$$
\begin{aligned}
R \;&=\; \left\{ \{s \in S \mid (w_s \cdot f_s)(z) \geq r\} \mid z \in \mathbb{R}^{4k},\ r \in [0,\infty) \right\} \\
&=\; \left\{ \{s \in S \mid w_s \cdot \min\{d_H^2(s, s_1(z))\dots, d_H^2(s, s_k(z))\} \geq r\} \mid z \in \mathbb{R}^{4k},\ r \in [0,\infty) \right\} \\
&=\; \left\{ \{s \in S \mid \forall i \in [k]:\ \sqrt{w_s} \cdot d_H(s, s_i(z)) \geq \sqrt{r}\} \mid z \in \mathbb{R}^{4k},\ r \in [0,\infty) \right\}.
\end{aligned}
$$

Setting $w_s' = \sqrt{w_s}$ for each $s \in S$ and $r' = \sqrt{r}$, we get that the ranges are

$$R \;=\; \left\{ \{s \in S \mid \forall i \in [k]:\ w_s' \cdot d_H(s, s_i(z)) \geq r'\} \mid z \in \mathbb{R}^{4k},\ r' \in [0,\infty) \right\}.$$

For each segment $s \in S$, consider the hypersurface λ_s in $\mathbb{R}^{4k+1}$ given by the graph of the function $z \in \mathbb{R}^{4k} \mapsto w_s' \cdot d_H(s, s_i(z))$. This is $\lambda_s \equiv \left\{ \left(z, w_s' \cdot d_H(s, s_i(z))\right) \in \mathbb{R}^{4k} \times \mathbb{R} \mid z \in \mathbb{R}^{4k} \right\}$. As it has been discussed and used in Section 3 when defining the set $\mathcal{F}(ab, a'b')$, the hypersurface λ_s is contained in the union of 8 *algebraic* hypersurfaces of bounded degree, each of them being the graph of a function. Let Λ_s be the set of those 8 algebraic hypersurfaces for the segment $s \in S$.

Set $\Lambda := \cup_{s\in S}\Lambda_s$ and let $\mathcal{A}$ be the arrangement in $\mathbb{R}^{4k+1}$ induced by Λ. Each point $(z, r') \in \mathbb{R}^{4k} \times \mathbb{R}$ gives a range to R, and two points in the same cell of $\mathcal{A}$ give exactly the

same range to R because, for each $s \in S$, the surface λ_s is above, below or on all the points of the cell. It may happen that points in different cells of $\mathcal{A}$ give the same range, as one still has to check the condition $\forall i \in [k] : w'_s \cdot d_H(s, s_i(z)) \geq r'$. In any case, the number of cells in $\mathcal{A}$ is an upper bound to the number of ranges in R, which is exactly the number of ranges in $\mathrm{range}_{\geq}(F_w)$.

Classical results in Real Algebraic Geometry imply that $\mathcal{A}$ has $|\Lambda|^{O(k)}$ cells; see for example [4, Chapter 7] or [32, Section 6.2]. This implies that the so-called shattering dimension of $\mathrm{range}_{\geq}(F_w)$ is $O(k) = O(1)$. (See for example Har-Peled [27, Chapter 5] for the concept and the next property.) Since a range space has bounded shattering dimension if and only if it has bounded VC-dimension this implies that the VC-dimension of $\mathrm{range}_{\geq}(F_w)$ is $O(1)$. The approach we have used is essentially an application of the methodology discussed by Matoušek [32, Section 10.3].

Note that in this proof we have not tried to optimize the bound on the VC-dimension because we assume k is constant. Perhaps a better bound follows from adapting the result of Driemel at al. [20] to the case of weights. ◀

We can now apply Theorem 6 on F to obtain the coreset.

▶ **Theorem 9.** *Assume that k is a fixed positive integer. Let δ, ε be real values in $(0, 1/2)$. For any set S of n unweighted segments in the plane, we can compute in time $O(n \log(1/\delta))$ a subset $T \subseteq S$ of*

$$O\left(\varepsilon^{-2} \log \frac{1}{\delta}\right)$$

segments and weights $u_s > 0$ for each $s \in T$ such that, with probability at least $1 - \delta$:

$$\forall \text{ segments } s_1, \ldots, s_k : \quad |\mathrm{cost}_S(\{s_1, \ldots, s_k\}) - \mathrm{cost}_T(\{s_1, \ldots, s_k\})| \leq \varepsilon \cdot \mathrm{cost}_S(\{s_1, \ldots, s_k\}).$$

Proof. We first compute a bicriteria ($\alpha = O(1), \beta = O(1)$)-approximation for k-means on S by using the algorithm of Chen [9, Theorem A.4], which in turn is a modification of the algorithm by Indyk [29]. For a probability of error $\delta' = \delta/2$, the algorithm takes $O(n \log(1/\delta')) = O(n \log(1/\delta))$ time and succeeds with probability at least $1 - \delta'$ in finding a set $s_1, \ldots, s_{k'}$ of $k' = O(k)$ segments such that $\mathrm{cost}_S(\{s_1, \ldots, s_{k'}\}) \leq O(1) \cdot \mathrm{cost}_S(\{s_1, \ldots, s_k\})$. Note that the algorithm of Chen is for the discrete version of k-means, where the centers under consideration must be a subset of S. However, it is well-known that the triangle inequality implies that this is a factor 4 off for the continuous k-means version. This factor 4 is then subsumed by the $O(1)$ approximation factor.

Let $F = \{f_s \mid s \in S\}$. We use the bicriteria approximation for the sensitivity upper bounds $\tilde{\sigma}(f_s)$, for each $f_s \in F$, as defined in Lemma 7. As discussed after Lemma 7, the total sensitivity $\tilde{\Sigma}(F)$ is $O(1)$ and, by Lemma 8, the VC-dimension of $\mathrm{range}_{\geq}(F)$ is $O(1)$. The result then follows using Theorem 6 with probability of error $\delta' = \delta/2$. The size of the set $C \subseteq F$ selected by Theorem 6 is $O(\tilde{\Sigma}(F)\varepsilon^{-2}(\log \tilde{\Sigma}(F) + \log(1/\delta'))) = O(\varepsilon^{-2} \log(1/\delta))$, and each function $f \in C$ has a given weight $u_f > 0$. We set $T = \{s \in S \mid f_s \in C\}$ and, for each segment $s \in T$, we define the weight $w_s := u_{f_s}$.

With probability at least $1 - \delta'$ we have

$$\forall z \in \mathbb{R}^{4k} : \quad \left| \sum_{f \in F} f(z) - \sum_{f \in C} u_f \cdot f(z) \right| \leq \varepsilon \sum_{f \in F} f(z),$$

which can be rewritten as

$$\forall z \in \mathbb{R}^{4k} : \quad \left| \sum_{s \in S} f_s(z) - \sum_{s \in T} w_s \cdot f_s(z) \right| \leq \varepsilon \sum_{s \in S} f_s(z).$$

Since $f_s(z) = \min\{d_H^2(s, s_1(z)), \ldots, d_H^2(s, s_k(z))\}$ and $s_1(z), \ldots, s_k(z)$ goes over all k tuples of candidate segments when z iterates over all $\mathbb{R}^{4k}$, the last statement is equivalent to

$$\forall \text{ segments } s_1, \ldots, s_k : \quad |\text{cost}_S(\{s_1, \ldots, s_k\}) - \text{cost}_T(\{s_1, \ldots, s_k\})| \leq \varepsilon \cdot \text{cost}_S(\{s_1, \ldots, s_k\}).$$

The algorithm may fail only if the bicriteria approximation of Chen fails or if the application of Theorem 6 fails, and each of them separately fails with probability at most $\delta/2$.   ◄

5   Putting it all together

Let S be a set of n segments in the plane without weights. We first set a fixed probability of error $\delta = 1/2$, which means that the terms $\log(1/\delta)$ become $O(1)$. We keep using $\varepsilon \in (0, 1/2)$ as a parameter.

We first compute a weighted coreset $T \subseteq S$ with $|T| = O(\varepsilon^{-2})$ elements in $O(n)$ time as described in Theorem 9; for each segment $s \in T$ we have a weight $w_s > 0$. If $S^* = \{s_1^*, \ldots, s_k^*\}$ is an optimal set of segments for S, then from Theorem 9 we have that $\text{cost}_T(S^*) \leq (1 + \varepsilon) \cdot \text{cost}_S(S^*)$ with probability at least $1/2$.

We apply the $(1 + \varepsilon)$-approximation algorithm of Theorem 5 on T, taking into account the weights of the segments. As $|T| = O(\varepsilon^{-2})$ the algorithm runs in time

$$O\left((\varepsilon^{-2})^{8k-2+\eta} + (\varepsilon^{-3})^{4k+1} \log^{4k+1}(\varepsilon^{-3})\right) = O\left(\varepsilon^{-16k+4-\eta} + \varepsilon^{-12k-3} \log^{4k+1}(\varepsilon^{-1})\right)$$

for any $\eta > 0$. When $k = 1$, the second summand dominates. When $k \geq 2$ and ε is below some constant ε_0, the first summand dominates.

Let $T^* = \{t_1^*, \ldots, t_k^*\}$ be an optimal set of segments for k-means of the weighted set T. The algorithm of Theorem 5 has then provided a set $S_\varepsilon = \{s_{1,\varepsilon}, \ldots, s_{k,\varepsilon}\}$ of k segments for which $\text{cost}_T(S_\varepsilon) \leq (1 + \varepsilon) \cdot \text{cost}_T(T^*)$. Note that for the set S_ε we also get from Theorem 9 that $(1 - \varepsilon) \cdot \text{cost}_S(S_\varepsilon) \leq \text{cost}_T(S_\varepsilon)$. Since $\text{cost}_T(T^*) \leq \text{cost}_T(S^*)$, we conclude that

$$\begin{aligned} (1 - \varepsilon) \cdot \text{cost}_S(S_\varepsilon) \; &\leq \; \text{cost}_T(S_\varepsilon) \; \leq \; (1 + \varepsilon) \cdot \text{cost}_T(T^*) \; \leq \; (1 + \varepsilon) \cdot \text{cost}_T(S^*) \\ &\leq \; (1 + \varepsilon)^2 \cdot \text{cost}_S(S^*) \end{aligned}$$

or

$$\text{cost}_S(S_\varepsilon) \; \leq \; \frac{(1 + \varepsilon)^2}{(1 - \varepsilon)} \cdot \text{cost}_S(S^*) \; = \; (1 + O(\varepsilon)) \cdot \text{cost}_S(S^*).$$

Setting $\varepsilon = \Theta(\varepsilon')$ appropriately, we get a $(1 + \varepsilon')$-approximation for any desired ε'.

By independently repeating the algorithm $O(\log(1/\delta))$ times and taking the best among the solutions, we can reduce the probability of error to any given value δ. Because $k = O(1)$, evaluating each candidate solution with respect to the whole set of segments takes $O(n)$ time. We summarize in the following.

▶ **Theorem 10.** *Let k a fixed, positive integer and let $\delta, \varepsilon \in (0, 1/2)$. Let S be a family of n unweighted segments in the plane. We can compute k segments $s_{1,\varepsilon}, \ldots, s_{k,\varepsilon}$ in $\mathbb{R}^2$ such that, with probability at least $1 - \delta$,*

$$\text{cost}_S(\{s_{1,\varepsilon}, \ldots, s_{k,\varepsilon}\}) \; \leq \; (1 + \varepsilon) \min_{s_1, \ldots, s_k} \text{cost}_S(\{s_1, \ldots, s_k\})$$

in time $O\left(\left(n + \varepsilon^{-16k+4-\eta} + \varepsilon^{-12k-3} \log^{4k+1}(\varepsilon^{-1})\right)(\log(1/\delta))\right)$, for any $\eta > 0$.

For $k = 1$, the running time is $O\left(\left(n + \varepsilon^{-15}\log^5(\varepsilon^{-1})\right)(\log(1/\delta))\right)$, while for $k \geq 2$ the running time is $O\left(\left(n + \varepsilon^{-16k+4-\eta}\right)(\log(1/\delta))\right)$ for any $\eta > 0$.

6 Extension to polylines

In this section we briefly discuss the extension of our result to the case of polylines of bounded complexity. To reduce the number of parameters, we assume that each polyline has at most ℓ segments and we search the k-means among polylines that have at most ℓ segments. (We can also handle the case where the input and the target centers have different complexities.) To simplify the discussion, we assume that each input polyline has exactly ℓ segments. We further assume that $\ell = O(1)$.

We regard each polyline π as the union of segments and note that the distance between the polyline π with segments $s_1, \ldots, s_\ell$ and the polyline π' with segments $s'_1, \ldots, s'_\ell$ is

$$d_H(\pi, \pi') \; = \; \max\left\{\max_{i\in[\ell]}\min_{j\in[\ell]} d_H(s_i, s'_j), \; \max_{j\in[\ell]}\min_{i\in[\ell]} d_H(s'_j, s_i)\right\}.$$

Therefore the distance between any two polylines is described as a max-min combination of $O(\ell^2) = O(1)$ values.

A polyline with ℓ segments is parameterized by $2(\ell+1)$ real values. Therefore, a sequence of k polylines with ℓ segments each is parameterized by a point in $\mathbb{R}^\kappa$ for $\kappa = 2k(\ell+1)$. (Before, for segments, we had $\kappa = 4k$.) Each $z \in \mathbb{R}^\kappa$ defines k polylines $\pi_1(z), \ldots, \pi_k(z)$, each consisting of ℓ segments.

Let Π be a set of polylines in the plane, each with ℓ segments. For each $\pi \in \Pi$, we define the function $f_\pi : \mathbb{R}^\kappa \to \mathbb{R}$ by

$$f_\pi(z) \; := \; \min\{d_H^2(\pi, \pi_1(z)), \ldots, d_H^2(\pi, \pi_k(z))\} \; = \; \left(\min\{d_H(\pi, \pi_1(z)), \ldots, d_H(\pi, \pi_k(z))\}\right)^2$$

and then define the set of functions $F = \{f_\pi \mid \pi \in \Pi\}$.

We first note that the VC-dimension of the range space $\text{range}_\geq(F_w)$ is $O(1)$, where F_w is obtained from F by scaling each $f_\pi \in F$ with a different scalar $w_\pi > 0$. The proof of Lemma 8 readily applies to this case as it only relies on the description complexity of $d_H(\pi, \pi_i(z))$ being constant, and each patch of the description being an algebraic function.

Next we note that we can use the bicriteria $(\alpha = O(1), \beta = O(1))$-approximation for k-means of Chen, as we did in the proof of Theorem 9. Indeed, this algorithm only requires that we can compute the distance between any two input objects, which we can do in constant time. The rest of the proof of Theorem 9 goes unchanged because Lemma 7 and Theorem 6 do not make any assumption related to segments beyond the VC-dimension. We thus obtain with probability at least $1/2$ a coreset $\tilde{\Pi}$ of $O(\varepsilon^{-2})$ input polylines, each of them with a positive weight w_π.

It remains to adapt Theorem 5 to the setting of polylines. As we have done in Theorem 4, for each polyline π we can compute a family $\mathcal{F}_\pi$ of nice functions such that

$$f_\pi(z) \; = \; \sum_{f \in \mathcal{F}_\pi} f(z) \; = \; \min_{i\in[k]} d_H^2(\pi, \pi_i(z)) \quad \text{for all } z \in \mathbb{R}^\kappa.$$

Indeed, as we did in the proof of Theorem 4, we can break the parameter space $\mathbb{R}^\kappa$ using $O(k^2\ell^2) = O(1)$ algebraic hypersurfaces into $O(1)$ cells such that, within each cell, the max-max-min expression defining $d_H(\pi, \pi_i(z))$ is always the same algebraic expression. We can then apply Theorem 3 to the family of nice functions $\cup_{\pi\in\tilde{\Pi}}\mathcal{F}_\pi$, where each function

in $\mathcal{F}_\pi$ has been scales with the corresponding weight w_π. Thus, we have an application of Theorem 3 in $\mathbb{R}^\kappa$ for $O(\varepsilon^{-2})$ functions. The running time is, for any $\eta > 0$,

$$O((\varepsilon^{-2})^{2\kappa-2+\eta}+(\varepsilon^{-3})^{\kappa+1}\log^{\kappa+1}(\varepsilon^{-3})) \;=\; O(\varepsilon^{-4\kappa+4+\eta}+\varepsilon^{-3\kappa-3}\log^{\kappa+1}(\varepsilon^{-}1)) \;=\; O(\varepsilon^{-O(k\ell)}).$$

Like before, we can make $O(\log(1/\delta))$ independent repetitions to decrease the probability of failure to δ. We summarize below.

▶ **Theorem 11.** *Let k and ℓ be fixed, positive integers and let $\delta, \varepsilon \in (0, 1/2)$ be parameters. Let Π be a family of n unweighted polylines in the plane, each with at most ℓ segments. We can compute k polylines $\pi_{1,\varepsilon}, \ldots, \pi_{k,\varepsilon}$ in $\mathbb{R}^2$, each with at most ℓ segments, such that, with probability at least $1 - \delta$,*

$$\mathrm{cost}_\Pi(\{\pi_{1,\varepsilon}, \ldots, \pi_{k,\varepsilon}\}) \;\leq\; (1+\varepsilon)\min_{\pi_1,\ldots,\pi_k}\mathrm{cost}_\Pi(\{\pi_1, \ldots, \pi_k\})$$

in time $O\left(\left(n + \varepsilon^{-O(k\ell)}\right)(\log(1/\delta))\right)$.

───── **References** ─────

1   Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, and Justin Ward. Better guarantees for k-means and Euclidean k-median by primal-dual algorithms. *SIAM J. Comput.*, 49(4), 2020.

2   Pranjal Awasthi, Moses Charikar, Ravishankar Krishnaswamy, and Ali Kemal Sinop. The hardness of approximation of Euclidean k-means. In *31st International Symposium on Computational Geometry, SoCG 2015*, volume 34 of *LIPIcs*, pages 754–767. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2015.

3   Sayan Bandyapadhyay and Kasturi R. Varadarajan. On variants of k-means clustering. In *32nd International Symposium on Computational Geometry, SoCG 2016*, volume 51 of *LIPIcs*, pages 14:1–14:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2016.

4   Saugata Basu, Richard Pollack, and Marie-Françoise Roy. *Algorithms in Real Algebraic Geometry.* Springer Berlin, Heidelberg, 2006.

5   Kevin Buchin, Anne Driemel, Joachim Gudmundsson, Michael Horton, Irina Kostitsyna, Maarten Löffler, and Martijn Struijs. Approximating (k, ℓ)-center clustering for curves. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019*, pages 2922–2938. SIAM, 2019.

6   Maike Buchin, Anne Driemel, and Dennis Rohde. Approximating (k, ℓ)-median clustering for polygonal curves. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021*, pages 2697–2717, 2021.

7   Frédéric Cazals, Bernard Delmas, and Timothee O'Donnell. Fréchet mean and p-mean on the unit circle: Decidability, algorithm, and applications to clustering on the flat torus. In *19th International Symposium on Experimental Algorithms, SEA 2021*, volume 190 of *LIPIcs*, pages 15:1–15:16, 2021.

8   Deeparnab Chakrabarty, Maryam Negahbani, and Ankita Sarkar. Approximation algorithms for continuous clustering and facility location problems. In *30th Annual European Symposium on Algorithms, ESA 2022*, volume 244 of *LIPIcs*, pages 33:1–33:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.

9   Ke Chen. On coresets for k-median and k-means clustering in metric and Euclidean spaces and their applications. *SIAM J. Comput.*, 39(3):923–947, 2009.

10   Siu-Wing Cheng and Haoqiang Huang. Curve simplification and clustering under Fréchet distance. In *Proceedings of the 2023 ACM-SIAM Symposium on Discrete Algorithms, SODA 2023*, pages 1414–1432. SIAM, 2023.

11   Vincent Cohen-Addad. A fast approximation scheme for low-dimensional k-means. In Artur Czumaj, editor, *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018*, pages 430–440. SIAM, 2018.

12 Vincent Cohen-Addad, Hossein Esfandiari, Vahab S. Mirrokni, and Shyam Narayanan. Improved approximations for Euclidean k-means and k-median, via nested quasi-independent sets. In Stefano Leonardi and Anupam Gupta, editors, *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1621–1628. ACM, 2022.

13 Vincent Cohen-Addad, Andreas Emil Feldmann, and David Saulpic. Near-linear time approximation schemes for clustering in doubling metrics. *J. ACM*, 68(6):44:1–44:34, 2021.

14 Vincent Cohen-Addad, Anupam Gupta, Amit Kumar, Euiwoong Lee, and Jason Li. Tight FPT approximations for k-median and k-means. In *46th International Colloquium on Automata, Languages, and Programming, ICALP 2019*, volume 132 of *LIPIcs*, pages 42:1–42:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019.

15 Vincent Cohen-Addad, Karthik C. S., and Euiwoong Lee. On approximability of clustering problems without candidate centers. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021*, pages 2635–2648, 2021.

16 Vincent Cohen-Addad, Karthik C. S., and Euiwoong Lee. Johnson coverage hypothesis: Inapproximability of k-means and k-median in p-metrics. In *Proceedings of the 2022 ACM-SIAM Symposium on Discrete Algorithms, SODA 2022*, pages 1493–1530. SIAM, 2022.

17 Vincent Cohen-Addad, Philip N. Klein, and Claire Mathieu. Local search yields approximation schemes for k-means and k-median in Euclidean and minor-free metrics. *SIAM Journal on Computing*, 48(2):644–667, 2019.

18 Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. Improved coresets and sublinear algorithms for power means in Euclidean spaces. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pages 21085–21098, 2021.

19 Anne Driemel, Amer Krivosija, and Christian Sohler. Clustering time series under the Fréchet distance. In Robert Krauthgamer, editor, *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016*, pages 766–785. SIAM, 2016.

20 Anne Driemel, André Nusser, Jeff M. Phillips, and Ioannis Psarros. The VC dimension of metric balls under Fréchet and Hausdorff distances. *Discret. Comput. Geom.*, 66(4):1351–1381, 2021.

21 Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011*, pages 569–578. ACM, 2011.

22 Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, PCA, and projective clustering. *SIAM Journal on Computing*, 49(3):601–657, 2020.

23 Daniel Ferguson and François G. Meyer. Computation of the sample Fréchet mean for sets of large graphs with applications to regression. In *Proceedings of the 2022 SIAM International Conference on Data Mining, SDM 2022*, pages 379–387, 2022.

24 Maurice Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de L'Institut Henri Poincaré*, 10(4):215–310, 1948.

25 Fabrizio Grandoni, Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Rakesh Venkat. A refined approximation for Euclidean k-means. *Inf. Process. Lett.*, 176:106251, 2022.

26 Sudipto Guha and Samir Khuller. Greedy strikes back: Improved facility location algorithms. *J. Algorithms*, 31(1):228–248, 1999.

27 Sariel Har-peled. *Geometric Approximation Algorithms*. American Mathematical Society, 2011.

28 Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 291–300. ACM, 2004.

29 Piotr Indyk. Sublinear time algorithms for metric space problems. In *Proceedings of the 31st Annual ACM Symposium on Theory of Computing*, pages 428–434. ACM, 1999.

30   Eric D. Kolaczyk, Lizhen Lin, Steven J. Rosenberg, Jie Xu, and Jackson Walters. Averages of unlabeled networks: Geometric characterization and asymptotic behavior. *The Annals of Statistics*, 48(1):514–538, 2020.

31   Amit Kumar, Yogish Sabharwal, and Sandeep Sen. Linear-time approximation schemes for clustering problems in any dimensions. *J. ACM*, 57(2):5:1–5:32, 2010.

32   Jiří Matoušek. *Lectures on Discrete Geometry*, volume 212 of *Graduate texts in mathematics*. Springer, 2002.

33   Nimrod Megiddo and Kenneth J. Supowit. On the complexity of some common geometric location problems. *SIAM J. Comput.*, 13(1):182–196, 1984.

34   François G. Meyer. The Fréchet mean of inhomogeneous random graphs. In *Complex Networks & Their Applications X - Volume 1, Proceedings of the Tenth International Conference on Complex Networks and Their Applications, COMPLEX NETWORKS 2021*, volume 1015 of *Studies in Computational Intelligence*, pages 207–219, 2021.

35   Yuriy Mileyko, Sayan Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007, 2011.

36   Abhinandan Nath and Erin Taylor. k-median clustering under discrete Fréchet and Hausdorff distances. *J. Comput. Geom.*, 12:156–182, 2022.

37   Christof Schötz. *The Fréchet Mean and Statistics in Non-Euclidean Spaces*. PhD thesis, Heidelberg University, The Faculty of Mathematics and Computer Science, 2021.

38   Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. Fréchet means for distributions of persistence diagrams. *Discret. Comput. Geom.*, 52(1):44–70, 2014.

39   Antoine Vigneron. Geometric optimization and sums of algebraic functions. *ACM Trans. Algorithms*, 10(1):4:1–4:20, 2014.