CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Dagstuhl Research Online Publication Server

Sketched MinDist

Jeff M. Phillips
School of Computing, University of Utah, Salt Lake City, UT, USA
http://www.cs.utah.edu/~jeffp/
jeffp@cs.utah.edu

Pingfan Tang
School of Computing, University of Utah, Salt Lake City, UT, USA
https://my.eng.utah.edu/~pingfant/
tang1984@cs.utah.edu

Abstract

We sketch geometric objects J as vectors through the MinDist function, setting the ith coordinate

$$v_i(J) = \inf_{p \in J} \|p - q_i\|$$

for $q_i \in Q$ from a point set Q. Building a vector from these coordinate values induces a simple, effective, and powerful distance: the Euclidean distance between these sketch vectors. This paper shows how large this set Q needs to be under a variety of shapes and scenarios. For hyperplanes we provide direct connection to the sensitivity sampling framework, so relative error can be preserved in d dimensions using $|Q| = O(d/\varepsilon^2)$. However, for other shapes, we show we need to enforce a minimum distance parameter ρ, and a domain size L. For $d = 2$ the sample size Q then can be $\tilde{O}((L/\rho) \cdot 1/\varepsilon^2)$. For objects (e.g., trajectories) with at most k pieces this can provide stronger *for all* approximations with $\tilde{O}((L/\rho) \cdot k^3/\varepsilon^2)$ points. Moreover, with similar size bounds and restrictions, such trajectories can be reconstructed exactly using only these sketch vectors. Cumulatively, these results demonstrate that these MinDist sketch vectors provide an effective and efficient shape model, a compact representation, and a precise representation of geometric objects.

2012 ACM Subject Classification Theory of computation → Computational geometry

Keywords and phrases curve similarity, sensitivity sampling, sketching

Digital Object Identifier 10.4230/LIPIcs.SoCG.2020.63

Related Version available at https://arxiv.org/abs/1907.02171.

Funding *Jeff M. Phillips*: Thanks to NSF CCF-1350888, ACI-1443046, CNS- 1514520, CNS-1564287, and IIS-1816149.

1 Introduction

In this paper we analyze a new sketch for geometric objects, which we introduced in a recent more empirically-focused paper [23]. For an object $J \in \mathcal{J}$, where $J \subset \mathbb{R}^d$, this depends on a set of *landmarks* $Q \subset \mathbb{R}^d$; for now let $n = |Q|$. These landmarks induce a *sketched representation* $v_Q(J) \in \mathbb{R}^n$ where the ith coordinate $v_i(J)$ is defined via a MinDist operation

$$v_i(J) = \mathsf{dist}(q_i, J) = \inf_{p \in J} \|p - q_i\|,$$

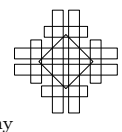
using the ith landmark $q_i \in Q$. When the object J is implicit, we simply use v_i. The most useful implication of this sketch is a simple new distance $\mathtt{d}_Q$ between two objects $J_1, J_2 \in \mathcal{J}$; the Euclidean distance between the (normalized as $\bar{v}_Q = \frac{1}{\sqrt{|Q|}} v_Q$) sketched representations

$$\mathtt{d}_Q(J_1, J_2) = \left\|\bar{v}_Q(J_1) - \bar{v}_Q(J_2)\right\|.$$

A second implication we will show, is that shapes J can often be recovered exactly from the sketch $v_Q(J)$ – demonstrating the richness of information it captures.

© Jeff M. Phillips and Pingfan Tang;
licensed under Creative Commons License CC-BY
36th International Symposium on Computational Geometry (SoCG 2020).
Editors: Sergio Cabello and Danny Z. Chen; Article No. 63; pp. 63:1–63:16

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



Our recent paper [23] introduces other variants of this distance (using other norms or using the $\arg\min_{p \in J}$ points on each $J \in \mathcal{J}$). We focus on this version as it is the simplest, cleanest, easiest to use, and was the best or competitive with the best on all empirical tasks. Indeed, for the pressing case of measuring a distance between trajectories, this new distance measure dominates a dozen other distance measures (including dynamic time warping, discrete Frechet distance, edit distance for real sequences) in terms of classification performance. In practice we find we only need $|Q| = 20$ landmarks to achieve high classification accuracy. It is also considerably more efficient in clustering and nearest neighbor tasks [23]; since it uses Euclidean distance, Lloyds algorithms works for k-means clustering and extremely efficient nearest neighbor packages [1, 25] automatically work with no extra engineering.

The goal of this paper is to formally understand how many landmarks in Q are needed for various error guarantees, and how to chose the locations of these points Q.

Our aims in the choice of Q are two-fold: first, we would like to approximate $\mathtt{d}_Q$ with $\mathtt{d}_{\tilde{Q}}$, and second we would like to recover $J \in \mathcal{J}$ exactly only using $v_Q(J)$. The specific results vary depending on the initial set Q and the object class $\mathcal{J}$. More precisely, the approximation goal aims to preserve d_Q for all objects J in some class $\mathcal{J}$ with a subset $\tilde{Q} \subset Q$ of landmarks. Or possibly a weighted set of landmarks $W, \tilde{Q}$ with $|\tilde{Q}| = N$, so each q_i is associated with a weight w_i and the weighted distance is defined

$$\mathtt{d}_{\tilde{Q},W}(J_1, J_2) = \sqrt{\sum_{i=1}^{N} w_i \cdot (v_i(J_1) - v_i(J_2))^2} = \left\| \tilde{v}_{\tilde{Q}}(J_1) - \tilde{v}_{\tilde{Q}}(J_2) \right\|$$

where $\tilde{v}_{\tilde{Q}} = (\tilde{v}_1, \cdots, \tilde{v}_N)$ with $\tilde{v}_i = \sqrt{w_i} v_i$. The set Q could also represent a continuous measure ω, which replaces w, and an integral on domain Ω replaces the sum. Specifically, our aim is an $(\rho, \varepsilon, \delta)$-*approximation of* Q *over* $\mathcal{J}$ so when $W, \tilde{Q}$ is selected by a random process that succeeds with probability at least $1 - \delta$, then for a *pair* $J_1, J_2 \in \mathcal{J}$ with $\mathtt{d}_Q(J_1, J_2) \geq \rho$

$$(1 - \varepsilon)\mathtt{d}_Q(J_1, J_2) \leq \mathtt{d}_{\tilde{Q},W}(J_1, J_2) \leq (1 + \varepsilon)\mathtt{d}_Q(J_1, J_2).$$

When this holds *for all* pairs in $\mathcal{J}$, we say it is a *strong* $(\rho, \varepsilon, \delta)$-*approximation of* Q *over* $\mathcal{J}$. In some cases we set to 0 either δ (the process is deterministic) or ρ (this preserves arbitrarily small distances), and may be able to use uniform weights $w_i = \frac{1}{|\tilde{Q}|}$ for all selected points.

1.1 Our Results

We begin with a special signed variant of the distance associated with the class $\mathcal{J}$ of $(d-1)$-dimensional hyperplanes (which for instance could model linear separators or linear regression models). This has $v_i(J)$ as negative on one side of the separator. In this variant, we show that if Q is full rank, then we can recover J from $v_Q(J)$, and a variant of sensitivity sampling can be used to select $O(d/(\delta\varepsilon^2))$ points to provide a $(0, \varepsilon, \delta)$-approximation $W, \tilde{Q}$. Or by selecting $O(\frac{d}{\varepsilon^2}(d \log d + \log \frac{1}{\delta}))$ results in a strong $(0, \varepsilon, \delta)$-approximation (Theorem 2).

Next we consider the more general case where the objects are bounded geometric objects $\mathcal{S}$. For such objects it is useful to consider a bounded domain $\Omega_L = [0, L]^d$ (for d a fixed constant), and consider the case where each $S \in \mathcal{S}$ and landmarks satisfy $S, Q \subset \Omega_L$. In this case, the number of samples required for a $(\rho, \varepsilon, \delta)$-approximation is $\mathfrak{S}_Q \frac{1}{\varepsilon^2 \delta}$ where

$$\mathfrak{S}_Q = O\left(\left(\frac{L}{\rho}\right)^{\frac{2d}{2+d}} \min\left(\log \frac{L}{\eta}, \log n, \left(\frac{L}{\rho}\right)^2 \right)^{\frac{2}{2+d}} \right), \tag{1}$$

where $\eta = \min_{q,q' \in Q} \|q - q'\|_\infty$. A few special cases are worth expanding upon. When Q is continuous and uniform over Ω_L then $\mathfrak{S}_Q = O((L/\rho)^{\frac{2d}{2+d}})$, and this is tight in $\mathbb{R}^2$ at $\mathfrak{S}_Q = \Theta(L/\rho)$. That is, we can show that $\mathfrak{S}_Q = \Theta(L/\rho)$ may be needed in general. When $d = 2$ but not necessarily uniform on Ω_L, then $\mathfrak{S}_Q = O(\frac{L}{\rho} \min\{\sqrt{\log n}, L/\rho\})$. And when Q is on a grid over Ω_L in $\mathbb{R}^2$ of resolution $\Theta(\rho)$, then $\mathfrak{S}_Q = O(\frac{L}{\rho}\sqrt{\log \frac{L}{\rho}})$, just a $\sqrt{\log L/\rho}$ factor more than the lower bound.

We conclude with some specific results for trajectories, represented as piecewise-linear curves. When considering the class $\mathcal{T}_k$ with at most k segments, then $O(\frac{1}{\varepsilon^2}\mathfrak{S}_Q(k^3 \log \mathfrak{S}_Q + \log \frac{1}{\delta}))$ samples is sufficient for a *strong* $(\rho, \varepsilon, \delta)$-approximation.

Also when considering trajectories $\mathcal{T}_\tau$ where the critical points are at distance at least τ apart from any non-adjacent part of the curve, we can *exactly reconstruct* the trajectory from v_Q as long as Q is a grid of side length $\Omega(\tau)$. It is much cleaner to describe the results for trajectories and Q precisely on a grid, but these results should extend for any object with k piecewise-linear boundaries, and critical points sufficiently separated, or Q as having any point in each sufficiently dense grid cell, as opposed to being exactly on the grid lattice.

1.2 Connections to other Domains, and Core Challenges

Before deriving these results, it is useful to lay out the connection to related techniques, including ones that our results will build on, and the challenges in applying them.

Sensitivity sampling. Sensitivity sampling [20, 16, 18, 26] is an important technique for our results. This typically considers a dataset X (a subset of a metric space), endowed with a measure $\mu : X \to \mathbb{R}^+$, and a family of cost functions F. These cost functions are usually related to the fitting of a data model or a shape S to X, and for instance on a single point $x \in X$, for $f \in F$, where

$$f(x) = \mathsf{dist}(x, S)^2 = \inf_{p \in S} \|x - p\|^2$$

is the squared distance from x to the closest point p on the shape S. And then $\bar{f} = \int_X f(x)\mathsf{d}\mu(x)$. The *sensitivity* [20] of $x \in X$ w.r.t. (F, X, μ) is defined as $\sigma_{F,X,\mu}(x) := \sup_{f \in F} \frac{f(x)}{\bar{f}}$, and the *total sensitivity* of F is defined as $\mathfrak{S}(F) = \int_X \sigma_{F,X,\mu}(x)\mathsf{d}\mu(x)$. This concept is quite general, and has been widely used in applications ranging from various forms of clustering [16, 18] to dimensionality reduction [17] to shape-fitting [26]. In particular, this will allow us to draw N samples $\tilde{X}$ iid from X proportional to $\sigma_{F,X,\mu}(x)$, and weighted $\tilde{w}(\tilde{x}) = \frac{\mathfrak{S}(F)}{N \cdot \sigma_{F,X,\mu}(\tilde{x})}$; we call this $\sigma_{F,X,\mu}$*-sensitive sampling*. Then $\tilde{X}$ is a $(0, \varepsilon, \delta)$-coreset; that is, with probability $1 - \delta$ for *each* $f \in F$

$$(1 - \varepsilon)\bar{f} \leq \int_{\tilde{X}} f(\tilde{x})\mathsf{d}\tilde{w}(\tilde{x}) \leq (1 + \varepsilon)\bar{f},$$

using $N = O(\frac{\mathfrak{S}(F)}{\varepsilon^2 \delta})$ [20]. The same error bound holds for *all* $f \in F$ (then it is called a $(0, \varepsilon, \delta)$*-strong coreset*) with $N = O(\frac{\mathfrak{S}(F)}{\varepsilon^2}(\mathfrak{s}_F \log \mathfrak{S}(F) + \log \frac{1}{\delta}))$ where $\mathfrak{s}_F$ is the shattering dimension of the range space $(X, \mathsf{ranges}(F))$ [5]. Each range $r \in (X, \mathsf{ranges}(F))$ is defined as points in a sublevel set of a cost function $r = \{x \in X \mid \frac{\mu(x)}{\mathfrak{S}(F)} \frac{f(x)}{\bar{f}} \leq \xi\}$ for some $f \in F$, $\xi \in \mathbb{R}$.

It seems natural that a form of our results would follow directly from these approaches. However, two significant and intertwined challenges remain. First, our goal is to approximate the distance between a pair of sketches $\|v_Q(J_1) - v_Q(J_2)\|$, whereas these results effectively only preserve the norm of a single sketch $\|v_Q(J_1)\|$; this prohibits many of the geometric

arguments in the prior work on this subject. Second, the total sensitivity $\mathfrak{S}(F)$ associated with unrestricted Q and pairs $J_1, J_2 \in \mathcal{J}$ is in general unbounded (as we prove in Section 3.1). Indeed, if the total sensitivity was bounded, it would imply a mapping to bounded vector space [20], wherein the subtraction of the two sketches $v_Q(J_1) - v_Q(J_2)$ would still be an element of this space, and the norm bound would be sufficient.

We circumvent these challenges in two ways. First, we identify a special case in Section 2 (with negative distances, for hyperplanes) under which there is a mapping of the sketch $v_Q(J_1)$ to metric space independent of the size and structure of Q. This induces a bound for total sensitivity related to a single object, and allows the subtraction of two sketches to be handled within the same framework.

Second, we enforce a lower bound on the distance $\mathtt{d}_Q(J_1, J_2) > \rho$ and an upper bound on the domain $\Omega_L = [0, L]^d$. This induces a restricted class of pairs $\mathcal{J}_{L/\rho}$ where L/ρ is a scaleless parameter, and it shows up in bounds we are then able to produce for the total sensitivity with respect to $\mathcal{J}_{L/\rho}$ and $Q \subset \Omega_L$.

Leverage scores, and large scales. The *leverage score* [14] of the ith column a_i of matrix A is defined as $\tau_i(A) := a_i^T (AA^T)^+ a_i$, where $(\cdot)^+$ is the Moore-Penrose pseudoinverse. This definition is more specific and linear-algebraic than sensitivity, but has received more attention for scalable algorithm development and approximation [14, 4, 13, 10, 22, 11].

However, the full version shows that if F is the collection of some functions defined on a set Q of n points ($\mu(q_i) = \frac{1}{n}$ for all $q_i \in Q$), where each $f \in F$ is the square of some function v in a finite dimensional space V spanned by a basis $\{v^{(1)}, \cdots, v^{(\kappa)}\}$, then we can build a $\kappa \times n$ matrix A where the ith column is $\frac{1}{\sqrt{n}}\big(v^{(1)}(q_i), \cdots, v^{(\kappa)}(q_i)\big)^T$, and then $\frac{1}{n} \cdot \sigma_{F,Q,\mu}(q_i)$ is precisely the leverage score of the ith column of the matrix A. A similar observation has been made by Varadarajan and Xiao [26].

A concrete implication of this connection is that we can invoke an online row sampling algorithm of Cohen *et al.* [11]. In our context, this algorithm would stream over Q, maintaining (ridge) estimates of the sensitivity of each q_i from a sample $\tilde{Q}_{i-1}$, and retaining each q_i in that sample based on this estimate. This provides a streaming approximation bound not much weaker than the sampling or gridding bounds we present; see full version.

MinDist and shape reconstruction. The fields of computational topology and surface modeling have extensively explored [6, 24, 8] the distance function to a compact set $J \subset \mathbb{R}^d$

$$\mathtt{d}_J(x) = \mathsf{dist}(x, J) = \inf_{p \in J} \|x - p\|,$$

their approximations, and the offsets $J^r = \mathtt{d}_J^{-1}([0, r])$. For instance the Hausdorff distance between two compact sets J, J' is $\mathtt{d}_{\mathsf{H}}(J, J') = \|\mathtt{d}_J - \mathtt{d}_{J'}\|_\infty$. The gradient of $\mathtt{d}_J$ implies stability properties about the medial axis [9]. And most notably, this stability of $\mathtt{d}_J$ with respect to a sample $P \sim J$ or $P \sim \partial J$ is closely tied to the development of shape reconstruction (aka geometric and topological inference) through α-shapes [15], power crust [2], and the like. The intuitive formulation through $\mathtt{d}_J$ (as opposed to Voronoi diagrams of P) has led to more statistically robust variants [8, 24] which also provide guarantees in shape recovery up to small feature size [7], essentially depending on the maximum curvature of ∂J.

Our formulation flips this around. Instead of considering samples P from J (or ∂J) we consider samples Q from some domain $\Omega \subset \mathbb{R}^d$. This leads to new but similar sampling theory, still depending on some feature size (represented by various scale parameters ρ, τ, and η), and still allowing recovery properties of the underlying objects. While the samples P

from J can be used to estimate Hausdorff distance via an all-pairs $O(|P|^2)$-time comparison, our formulation requires only a $O(|Q|)$-time comparison to compute $\mathtt{d}_Q$. We leave as open questions the recovering of topological information about an object $J \in \mathcal{J}$ from $v_Q(J)$.

2 The Distance Between Two Hyperplanes using Signed Sketches

A more detailed derivation of the results in this section are presented in the full version where proofs and a few technical details require more careful notation to navigate.

Let $\mathcal{H} = \{h \mid h \text{ is a hyperplane in } \mathbb{R}^d\}$ represent the space of all hyperplanes. Each hyperplane h can be represented by a vector $u \in \mathbb{R}^{d+1}$ composed as a normal vector $\bar{u} = (u_1, \ldots, u_d) \in \mathbb{R}^d$ with $\|\bar{u}\| = 1$ and offset u_{d+1}. Then the ith coordinate of a sketch vector can be derived as a *signed* distance from q_i as $v_i(h) = u_{d+1} + \langle \bar{u}, q_i \rangle$.

Recovery. Our recent paper [23] showed that if Q is full rank (there exist $d+1$ points in Q not on a common hyperplane) then $\mathtt{d}_Q(h_1, h_2) \neq 0$ if $h_1 \neq h_2$, and thus $\mathtt{d}_Q$ is a metric. This full rank condition on Q is also sufficient to recover h from $v_Q(h)$; e.g., using PCA.

Distance Preservation. Next we show that we can use a σ-sensitive sample $\tilde{Q}, W$ as a $(0, \varepsilon, \delta)$-coreset for this formulation; that is $\mathtt{d}_{\tilde{Q},W}$ preserves relative error with respect to $\mathtt{d}_Q$. We assume Q is full rank, and has uniform weight $\mu = \frac{1}{n}$ on each point. Using $X = Q$ we need to define the family of functions F to complete the tuple (F, Q, μ). To this end, let V be a $(d+1)$-dimensional function space with each element v_u is a linear function defined $v_u(q_i) = v_i(h) = u_{d+1} + \langle \bar{u}, q_i \rangle$. Now each $f \in F$ is defined as $f(q) = v(q)^2$, and through its representation u, each $h \in \mathcal{H}$ maps to a unique element of F.

However, we are interested in preserving $\mathtt{d}_Q$ which requires a pair $h_1, h_2 \in \mathcal{H}$, with corresponding normals $u^{(1)}, u^{(2)}$. Since V is a linear function space, then using $u = u^{(1)} - u^{(2)}$, then $f_{h_1,h_2}(q) = v_u(q)^2$, and $\mathtt{d}_Q(h_1, h_2) = (\frac{1}{n} \sum_{q \in Q} f_{h_1,h_2}(q))^{\frac{1}{2}}$. Note that u will likely not correspond to a single halfspace since the first d coordinates may not be a unit vector, but that is not an issue for this framework. Using Langberg and Schulman [20], the total sensitivity is $d+1$, and the sensitivities can be calculated (e.g., via leverage scores).

▶ **Theorem 1.** *Consider full rank $Q \subset \mathbb{R}^d$ and halfspaces $\mathcal{H}$ with $\varepsilon, \delta \in (0,1)$. A σ-sensitive sample $\tilde{Q}$ of (Q, F) of size $|\tilde{Q}| \geq \frac{d+1}{\delta \varepsilon^2}$ results in a $(0, \varepsilon, \delta)$-coreset. It is an $(0, \varepsilon, \delta)$-approximation so with probability at least $1 - \delta$, for each pair $h_1, h_2 \in \mathcal{H}$*

$$(1 - \varepsilon)\mathtt{d}_Q(h_1, h_2) \leq \mathtt{d}_{\tilde{Q},W}(h_1, h_2) \leq (1 + \varepsilon)\mathtt{d}_Q(h_1, h_2).$$

We can also achieve a strong coreset for this variant using results from Braverman *et al.* [5]. For this we need to provide an additional bound about the shattering dimension $\mathfrak{s} = \dim(Q, \mathcal{X})$ associated with each $f \in F$ and a weight $w : Q \to \mathbb{R}_+$. The range in the range space associated with f_{h_1,h_2} is defined for some η as

$$X_{h_1,h_2,\eta} = \{q \in Q \mid w(q) f_{h_1,h_2}(q) \leq \eta\}.$$

For $f \in F$ defined by a single halfspace, this is classically known to be $O(d)$. For the more general functions $f_{h_1,h_2} \in F$ defined by two halfspaces h_1, h_2, the same asymptotic bound can be shown using straight-forward decomposition properties of range spaces (see full version for proof). Then we can obtain the following result.

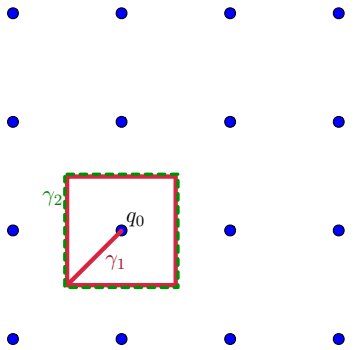


Figure 1 Q is the set of blue points, γ_1 is the red curve, γ_2 is the green (dashed) curve, and they coincide with each other on the boundary of the square.

▶ **Theorem 2.** *Consider full rank $Q \subset \mathbb{R}^d$ and halfspaces $\mathcal{H}$ with $\varepsilon, \delta \in (0,1)$. A σ-sensitive sample $\tilde{Q}$ of (Q,F) of size $|\tilde{Q}| = O(\frac{d}{\varepsilon^2}(d \log d + \log \frac{1}{\delta}))$ results in a strong $(0,\varepsilon,\delta)$-coreset. And thus a strong $(0,\varepsilon,\delta)$-approximation so with probability at least $1-\delta$, for all $h_1, h_2 \in \mathcal{H}$*

$$(1-\varepsilon)\mathtt{d}_Q(h_1,h_2) \leq \mathtt{d}_{\tilde{Q},W}(h_1,h_2) \leq (1+\varepsilon)\mathtt{d}_Q(h_1,h_2).$$

3 Sketched MinDist for Two Geometric Objects

In this section, we mildly restrict $\mathtt{d}_Q$ to the distance between any two geometric objects, in particular, bounded closed sets. Let $\mathcal{S} = \{S \subset \mathbb{R}^d \mid S \text{ is a bounded closed set}\}$ be the space of objects $\mathcal{J}$ we consider.

As before define $v_i(S) = \inf_{p \in S} \|p - q_i\|$, and then for $S_1, S_2 \in \mathcal{S}$ define $f_{S_1,S_2}(q_i) = (v_i(S_1) - v_i(S_2))^2$. The associated function space is $F(\mathcal{S}) = \{f_{S_1,S_2} \mid S_1, S_2 \in \mathcal{S}\}$. Setting $\mu(q) = \frac{1}{n}$ for all $q \in Q$, then $(\mathtt{d}_Q(S_1,S_2))^2 = \bar{f}_{S_1,S_2} := \sum_{i=1}^n \mu(q_i) f_{S_1,S_2}(q_i)$. Using sensitivity sampling to estimate $\mathtt{d}_Q(S_1,S_2)$ requires a bound on the total sensitivity of $F(\mathcal{S})$.

We show that while the total sensitivity $\mathfrak{S}(F(\mathcal{S}))$ is unbounded in general, it is tied to the ratio L/ρ between the diameter of the domain L, and the minimum allowed $\mathtt{d}_Q$ distance between objects ρ. In particular, it can be at least proportional to this, and in $\mathbb{R}^2$ in most cases (e.g., for near-uniform Q) is at most proportional to L/ρ or not much larger for any Q.

3.1 Lower Bound on Total Sensitivity

Suppose Q is a set of n points in $\mathbb{R}^2$ and no two points are at the same location, then for any $q_0 \in Q$ we can draw two curves γ_1, γ_2 as shown in Figure 1, where γ_1 is composed by five line segments and γ_2 is composed by four line segments. The four line segments of the γ_2 forms a square, on its boundary γ_1 and γ_2 coincide with each other, and inside this square, q_0 is the endpoint of γ_1. We can make this square small enough, such that all points $q \neq q_0$ are outside this square. So, we have $\mathsf{dist}(q_0,\gamma_1) = 0$ and $\mathsf{dist}(q_0,\gamma_2) \neq 0$, and $\mathsf{dist}(q,\gamma_1) = \mathsf{dist}(q,\gamma_2) = 0$ for all $q \neq q_0$. Thus, we have $f_{\gamma_1,\gamma_2}(q_0) > 0$ and $f_{\gamma_1,\gamma_2}(q) = 0$ for all $q \neq q_0$, which implies

$$\sigma_{F(\mathcal{S}),Q,\mu}(q_0) \geq \frac{f_{\gamma_1,\gamma_2}(q_0)}{\bar{f}_{\gamma_1,\gamma_2}} = \frac{f_{\gamma_1,\gamma_2}(q_0)}{\frac{1}{n}\sum_{q \in Q} f_{\gamma_1,\gamma_2}(q)} = \frac{n f_{\gamma_1,\gamma_2}(q_0)}{f_{\gamma_1,\gamma_2}(q_0)} = n.$$

Since this construction of two curves γ_1, γ_2 can be repeated around any point $q \in Q$,

$$\mathfrak{S}(F(\mathcal{S})) = \sum_{q \in Q} \mu(q)\sigma_{F(\mathcal{S}),Q,\mu}(q) \geq \sum_{q \in Q} \frac{1}{n} n = n.$$

We can refine this bound by introducing two parameters L, ρ for $\mathcal{S}$. Given $L > \rho > 0$ and a set $Q \subset \mathbb{R}^d$ of n points, we define $\mathcal{S}(L) = \{S \in \mathcal{S} \mid S \subset [0, L]^d\}$ and $F(\mathcal{S}(L), \rho) = \{f_{S_1,S_2} \in F(\mathcal{S}) \mid S_1, S_2 \in \mathcal{S}(L),\ \mathtt{d}_Q(S_1, S_2) \geq \rho\}$. The following lowerbounds the total sensitivity of $F(\mathcal{S}(L), \rho)$ for $d = 2$; it holds for any $d \geq 2$ using the construction in a 2d subspace.

▶ **Lemma 3.** *For $d = 2$ we can construct a set $Q \subset [0, L]^2$ such that $\mathfrak{S}(F(\mathcal{S}(L), \rho)) = \Omega(\frac{L}{\rho})$.*

Proof. We uniformly partition $[0, L]^2$ into n grid cells, such that $C_1 \frac{L}{\rho} \leq n \leq C_2 \frac{L}{\rho}$ for constants $C_1, C_2 \in (0, 1)$. The side length of each grid is $\eta = \frac{L}{\sqrt{n}}$. We take Q as the n grid points, and for each point $q \in Q$ we can choose two curves γ_1 and γ_2 (similar to curves in Figure 1) such that $\mathsf{dist}(q, \gamma_1) = 0$, $\mathsf{dist}(q, \gamma_2) \geq C_2\eta$, and $\mathsf{dist}(q', \gamma_1) = \mathsf{dist}(q', \gamma_2) = 0$ for all $q' \in Q \setminus \{q\}$. Thus, $\mathtt{d}_Q(\gamma_1, \gamma_2) \geq C_2 \frac{\eta}{\sqrt{n}} = C_2 \frac{L}{n} \geq \rho$. So, $f_{\gamma_1,\gamma_2} \in F(\mathcal{S}(L), \rho))$ and $\sigma(q) \geq n$ for all $q \in Q$ and $\mathfrak{S}(F(\mathcal{S}(L), \rho)) \geq n \geq C_1 \frac{L}{\rho}$, which implies $\mathfrak{S}(F(\mathcal{S}(L), \rho)) = \Omega(\frac{L}{\rho})$. ◀

3.2 Upper Bound on the Total Sensitivity

A simple upper bound of $\mathfrak{S}(F(\mathcal{S}(L), \rho)$ is $O(\frac{L^2}{\rho^2})$; it follows from the L/ρ constraint. The sensitivity of each point $q \in Q$ is defined as $\sup_{f_{S_1,S_2} \in F(\mathcal{S}(L),\rho)} \frac{f_{S_1,S_2}(q)}{\bar{f}_{S_1,S_2}}$, where $f_{S_1,S_2}(q) = O(L^2)$ for all $S_1, S_2 \in \mathcal{S}(L)$ and $q \in Q \subset [0, L]^d$, and the denominator $\bar{f}_{S_1,S_2} \geq \rho^2$ by assumption for all $f_{S_1,S_2} \in F(\mathcal{S}(L), \rho)$. Hence, the sensitivity of each point in Q is $O(\frac{L^2}{\rho^2})$, and thus their average, the total sensitivity is $O(\frac{L^2}{\rho^2})$. In this section we will improve and refine this bound.

We introduce two variables that only depends on $Q = \{q_1, \cdots, q_n\} \subset [0, L]^d$:

$$C_q := \max_{0<r\leq L} \frac{r^d}{L^d} \frac{n}{|Q \cap B_\infty(q, r)|} \quad \text{for } q \in Q, \text{ and } C_Q := \frac{1}{n} \sum_{q \in Q} C_q^{\frac{2}{2+d}}. \tag{2}$$

where $B_\infty(q, r) := \{x \in \mathbb{R}^d \mid \|x - q\|_\infty \leq r\}$. Intuitively, $\frac{|Q \cap B_\infty(q,r)|}{r^d}$ is proportional to the point density in region $B_\infty(q, r)$, and the value of $\frac{r^d}{L^d} \frac{n}{|Q \cap B_\infty(q,r)|}$ can be maximized, when the region $B_\infty(q, r)$ has smallest point density, which means r should be as large as possible but the number of points contained in $B_\infty(q, r)$ should be as small as possible. A trivial bound of C_q is n, but if we make $C_{q_0} = n$ for one point q_0, then it implies the value of C_q for other points will be small, so for C_Q it is possible to obtain a bound better than $n^{\frac{2}{d+2}}$.

Importantly, these quantities C_q and C_Q will be directly related to the sensitivity of a single point $\sigma(q)$ and the total sensitivity of the point set $\mathfrak{S}_Q$, respectively. We formalize this in two technical lemmas: First (in Lemma 7) $\sigma(q) \leq O((C_q(L/\rho)^d)^{\frac{2}{2+d}})$ and hence $\mathfrak{S}_Q = O(C_Q \cdot (L/\rho)^{\frac{2d}{2+d}})$; and second (in Lemma 8) we show $C_Q \leq O((\min\{\log \frac{L}{\eta}, \log n\})^{\frac{2}{2+d}})$ for Q of size n and $\eta = \min_{q,q' \in Q,\ q \neq q'} \|q - q'\|_\infty$.

Since $f_{S_1,S_2} \in F(\mathcal{S}(L), \rho)$, we know $f_{S_1,S_2}(q) \leq dL^2$ for all $q \in Q$ and $\frac{1}{n} \sum_{q' \in Q} f_{S_1,S_2}(q') \geq \rho^2$, so $\sigma(q) \leq \frac{dL^2}{\rho^2}$ for all $q \in Q$. Thus, we can expand $\frac{1}{|Q|} \sum_{q \in Q} \sigma(q)$ using Lemma 7 and factor out C_Q using Lemma 8 to immediately obtain the following theorem.

▶ **Theorem 4.** *Suppose $L > \rho > 0$, $Q = \{q_1, \cdots, q_n\} \subset [0, L]^d$ and $\eta = \min_{q,q' \in Q,\ q \neq q'} \|q - q'\|_\infty$. Then, we have*

$$\mathfrak{S}(F(\mathcal{S}(L), \rho)) \leq \mathfrak{S}_Q = O\left(\left(\frac{L}{\rho}\right)^{\frac{2d}{2+d}} \min\left(\log \frac{L}{\eta}, \log n, \left(\frac{L}{\rho}\right)^2\right)^{\frac{2}{2+d}}\right).$$

From Lemma 7 and Theorem 4, using [20][Lemma 2.1] we can obtain the following.

▶ **Theorem 5.** *Let $L > \rho > 0$, $Q = \{q_1, \cdots, q_n\} \subset [0, L]^d$, $S_1, S_2 \in \mathcal{S}(L)$ and $\mathtt{d}_Q(S_1, S_2) \geq \rho$. Then for $\delta, \varepsilon \in (0, 1)$ a σ-sensitive sampling of size $N \geq \frac{\mathfrak{S}_Q}{\delta\varepsilon^2}$ provides $\tilde{Q}$, a $(\rho, \varepsilon, \delta)$-coreset; that is with probability at least $1 - \delta$, we have*

$$(1 - \varepsilon)\mathtt{d}_Q(S_1, S_2) \leq \mathtt{d}_{\tilde{Q},W}(S_1, S_2) \leq (1 + \varepsilon)\mathtt{d}_Q(S_1, S_2).$$

If Q describes a continuous uniform distribution in $[0, L]^d$ (or sufficiently close to one, like points on a grid), then there exists an absolute constant $C > 0$ such that $C_q \leq C$ for all $q \in Q$, then in Lemma 7 $\sigma(q) \leq C_d\big(\frac{L}{\rho}\big)^{\frac{2d}{2+d}}$ for all $q \in Q$, and in Theorem 4 $\mathfrak{S}_Q \leq C_d\big(\frac{L}{\rho}\big)^{\frac{2d}{2+d}}$. So, for uniform distribution, the sample size of Q in Theorem 5 is independent from the size of Q, and for $d = 2$ the bound $\mathfrak{S}_Q = O(L/\rho)$ matches the lower bound in Lemma 3.

▶ **Corollary 6.** *If Q describes the continuous uniform distribution over $[0, L]^d$, then the sample size in Theorem 5 can be reduced to $N = O\Big(\big(\frac{L}{\rho}\big)^{\frac{2d}{2+d}} \frac{1}{\delta\varepsilon^2}\Big)$.*

Technical lemmas bounding $\sigma(q)$ and C_Q.

▶ **Lemma 7.** *For function family $F(\mathcal{S}(L), \rho)$ the sensitivity for any $q \in Q \subset [0, L]^d$ is bounded*

$$\sigma(q) \leq C_d C_q^{\frac{2}{2+d}} \left(\frac{L}{\rho}\right)^{\frac{2d}{2+d}},$$

where $C_d = 4^{\frac{2}{2+d}}(8\sqrt{d})^{\frac{2d}{2+d}}$ and C_q given by (2).

Proof. Recall $\sigma(q) = \sup_{f_{S_1,S_2} \in F(\mathcal{S}(L),\rho)} \frac{f_{S_1,S_2}(q)}{\frac{1}{n}\sum_{q' \in Q} f_{S_1,S_2}(q')}$. For any fixed $q \in Q$, for now suppose $f_{S_1,S_2} \in F(\mathcal{S}(L), \rho)$ satisfies this supremum $\sigma(q) = \frac{f_{S_1,S_2}(q)}{\frac{1}{n}\sum_{q' \in Q} f_{S_1,S_2}(q')}$. We define $\mathsf{dist}(q, S) = \inf_{p \in S} \|q - p\|$ (so for $q_i \in Q$ then $\mathsf{dist}(q_i, S) = v_i(S)$), and then use the parameter $M := |\mathsf{dist}(q, S_1) - \mathsf{dist}(q, S_2)|$, where $M^2 = f_{S_1,S_2}(q)$. If $M = 0$, then obviously $f_{S_1,S_2}(q) = M^2 = 0$, and $\sigma(q) = 0$. So, without loss of generality, we assume $M > 0$ and $\mathsf{dist}(q, S_1) = \tau$ and $\mathsf{dist}(q, S_2) = \tau + M$. We first prove $\sigma(q) \leq C_d C_q \frac{L^d}{M^d}$. There are two cases for the relationship between τ and M, as shown in Figure 2.

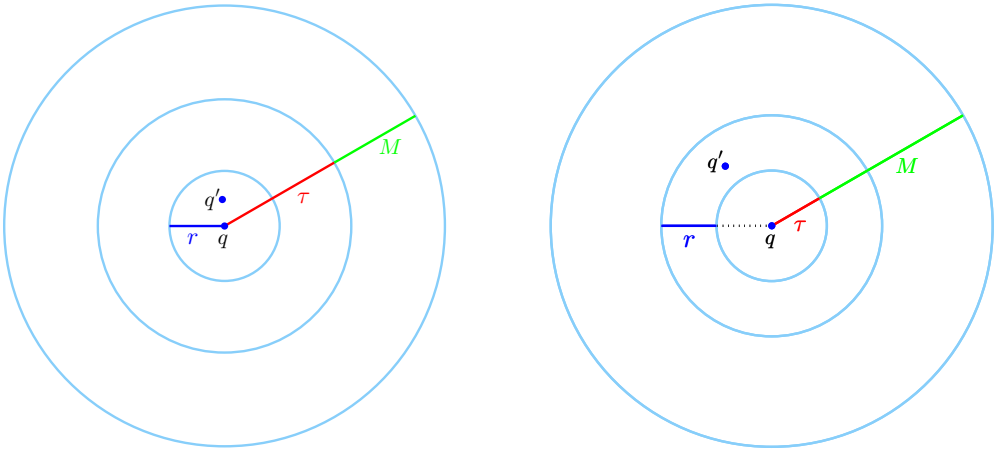


Figure 2 Left: Case 1, $r = \frac{M}{8} \leq \tau$, and $q' \in B(q, r)$. Right: Case 2, $r = \frac{M}{8} > \tau$, and $q' \in B(q, \tau + r)$.

Case 1: $\tau \geq \frac{M}{8}$. For any $q' \in B(q, \frac{M}{8}) := \{q' \in \mathbb{R}^d \mid \|q' - q_i\| \leq \frac{M}{8}\}$, we have $\tau + M = \mathsf{dist}(q, S_2) \leq \mathsf{dist}(q, q') + \mathsf{dist}(q', S_2) \leq \frac{M}{8} + \mathsf{dist}(q', S_2)$, which implies for all $q' \in B(q, \frac{M}{8})$

$$\mathsf{dist}(q', S_2) \geq \tau + M - \frac{M}{8} = \tau + \frac{7}{8}M.$$

Similarly $\mathsf{dist}(q', S_1) \leq \mathsf{dist}(q', q) + \mathsf{dist}(q, S_1) \leq \frac{M}{8} + \tau$ for all $q' \in B(q, \frac{M}{8})$. Thus for all $q' \in B(q, \frac{M}{8})$

$$|\mathsf{dist}(q', S_2) - \mathsf{dist}(q', S_1)| \geq \mathsf{dist}(q', S_2) - \mathsf{dist}(q', S_1) \geq \tau + \frac{7}{8}M - (\tau + \frac{M}{8}) = \frac{3}{4}M.$$

Case 2: $0 \leq \tau < \frac{M}{8}$. For any $q' \in B(q, \tau + \frac{M}{8}) := \{q' \in \mathbb{R}^d \mid \mathsf{dist}(q', q) \leq \tau + \frac{M}{8}\}$, we have $\tau + M = \mathsf{dist}(q', S_2) \leq \mathsf{dist}(q, q') + \mathsf{dist}(q', S_2) \leq \tau + \frac{M}{8} + \mathsf{dist}(q', S_2)$, which implies for all $q' \in B(q, \tau + \frac{M}{8})$

$$\mathsf{dist}(q', S_2) \geq \frac{7}{8}M.$$

Combined with $\tau < \frac{M}{8}$ and $\mathsf{dist}(q', S_1) \leq \mathsf{dist}(q', q) + \mathsf{dist}(q, S_1) \leq \tau + \frac{M}{8} + \tau = \frac{M}{8} + \frac{M}{8} + \frac{M}{8} \leq \frac{3}{8}M$ for all $q' \in B(q, \tau + \frac{M}{8})$, we have

$$|\mathsf{dist}(q', S_2) - \mathsf{dist}(q', S_1)| \geq \mathsf{dist}(q', S_2) - \mathsf{dist}(q', S_1) \geq \frac{7}{8}M - \frac{3}{8}M = \frac{M}{2}.$$

Combining these two cases on τ, for all $q' \in B(q, \frac{M}{8})$ $|\mathsf{dist}(q', S_2) - \mathsf{dist}(q', S_1)| \geq \frac{M}{2}$. Then since $B_\infty(q, \frac{r}{\sqrt{d}}) \subset B(q, r)$ for all $r \geq 0$, from

$$C_q = \max_{0<r\leq L} \frac{r^d}{L^d} \frac{n}{|Q \cap B_\infty(q, r)|} \geq (\frac{1}{8\sqrt{d}})^d \frac{M^d}{L^d} \frac{n}{|Q \cap B_\infty(q, \frac{M}{8\sqrt{d}})|},$$

we can bound the denominator in $\sigma(q)$ as

$$\begin{aligned}\frac{1}{n}\sum_{q'\in Q} f_{S_1,S_2}(q') &\geq \frac{1}{n} \sum_{q'\in Q\cap B_\infty(q,\frac{M}{8\sqrt{d}})} f_{S_1,S_2}(q') = \frac{1}{n} \sum_{q'\in Q\cap B_\infty(q,\frac{M}{8\sqrt{d}})} (\mathsf{dist}(q', S_1) - \mathsf{dist}(q', S_2))^2 \\ &\geq \frac{1}{4}\frac{1}{n}M^2 \left|Q \cap B_\infty(q, \frac{M}{8\sqrt{d}})\right| \geq \frac{1}{4}(\frac{1}{8\sqrt{d}})^d \frac{M^2}{C_q}\frac{M^d}{L^d} = \frac{1}{4}(\frac{1}{8\sqrt{d}})^d \frac{1}{C_q}\frac{M^{2+d}}{L^d},\end{aligned}$$

which implies

$$\sigma(q) = \frac{M^2}{\frac{1}{n}\sum_{q'\in Q} f_{S_1,S_2}(q')} \leq 4(8\sqrt{d})^d M^2 C_q \frac{L^d}{M^{2+d}} = 4(8\sqrt{d})^d C_q \frac{L^d}{M^d}.$$

Combining this with $\sigma(q) \leq \frac{M^2}{\rho^2}$, we have $\sigma(q) \leq \min\left(\frac{M^2}{\rho^2}, 4(8\sqrt{d})^d C_q \frac{L^d}{M^d}\right)$. If $M^{2+d} \leq 4(8\sqrt{d})^d C_q \rho^2 L^d$, then $\frac{M^2}{\rho^2} \leq 4(8\sqrt{d})^d C_q \frac{L^d}{M^d}$, which means $\sigma(q) \leq \min\left(\frac{M^2}{\rho^2}, 4(8\sqrt{d})^d C_q \frac{L^d}{M^d}\right) = \frac{M^2}{\rho^2} \leq 4^{\frac{2}{2+d}}(8\sqrt{d})^{\frac{2d}{2+d}} C_q^{\frac{2}{2+d}} \left(\frac{L}{\rho}\right)^{\frac{2d}{2+d}}$. If $M^{2+d} \geq 4(8\sqrt{d})^d C_q \rho^2 L^d$, then $4(8\sqrt{d})^d C_q \frac{L^d}{M^d} \leq \frac{M^2}{\rho^2}$, so $\sigma(q) \leq \min\left(\frac{M^2}{\rho^2}, 4(8\sqrt{d})^d C_q \frac{L^d}{M^d}\right) = 4(8\sqrt{d})^d C_q \frac{L^d}{M^d} \leq 4^{\frac{2}{2+d}}(8\sqrt{d})^{\frac{2d}{2+d}} C_q^{\frac{2}{2+d}} \left(\frac{L}{\rho}\right)^{\frac{2d}{2+d}}$. ◀

Hence, to bound the total sensitivity of $F(\mathcal{S}(L), \rho)$, we need a bound of $C_Q = \frac{1}{n}\sum_{q\in Q} C_q^{\frac{2}{2+d}}$.

▶ **Lemma 8.** *Suppose $Q \subset [0,L]^d$ of size n, $\eta = \min_{q,q' \in Q,\ q \neq q'} \|q - q'\|_\infty$, and C_Q is given by (2). Then using $C_d = 2^{d+1}$ we have*

$$C_Q \leq C_d \min\Big(\big(\log_2 \frac{L}{\eta}\big)^{\frac{2}{2+d}}, \big(\frac{1}{d}\log_2 n\big)^{\frac{2}{2+d}} \Big).$$

Proof. We define $\widetilde{C}_Q := \frac{1}{n}\sum_{q\in Q} C_q$, and using Hölder inequality we have

$$C_Q = \frac{1}{n}\sum_{q\in Q} C_q^{\frac{2}{2+d}} \leq \frac{1}{n}\Big(\sum_{q\in Q} C_q\Big)^{\frac{2}{2+d}} n^{\frac{d}{2+d}} = \Big(\frac{1}{n}\sum_{q\in Q} C_q\Big)^{\frac{2}{2+d}} = (\widetilde{C}_Q)^{\frac{2}{2+d}}.$$

So, we only need to bound $\widetilde{C}_Q$.

We define $r_q := \arg\max_{0<r\leq L} \frac{r^d}{L^d} \frac{n}{|Q \cap B_\infty(q,r)|}$ for all $q \in Q$, $Q_i := \{q \in Q \mid \frac{L}{2^{i+1}} < r_q \leq \frac{L}{2^i}\}$, and $A := \{i \geq 0 \mid i \text{ is an integer and } |Q_i| > 0\}$.

For any fixed $i \in A$, we use $l_i := \frac{L}{2^{i+1}}$ as the side length of grid cell to partition the region $[0,L]^d$ into $s_i = (\frac{L}{l_i})^d = 2^{(i+1)d}$ grid cells: $\Omega_1 . \cdots, \Omega_{s_i}$ where each Ω_j is a closed set, and define $Q_{i,j} := Q_i \cap \Omega_j$. Then, $|Q_i \cap \bar{B}_\infty(q,l_i)| \geq |Q_{i,j}|$ for all $q \in Q_{i,j}$ where $\bar{B}_\infty(q,l_i) := \{q' \in \mathbb{R}^d \mid \|q' - q\|_\infty \leq l_i\}$, and we have

$$\begin{aligned}
\sum_{q\in Q_i} \frac{r_q^d}{L^d} \frac{1}{|Q_i \cap B_\infty(q,r_q)|} &\leq \sum_{q\in Q} \frac{L^d}{2^{id}L^d} \frac{1}{|Q_i \cap B_\infty(q,r_q)|} \leq \frac{1}{2^{id}} \sum_{q\in Q_i} \frac{1}{|Q_i \cap \bar{B}_\infty(q,l_i)|} \\
&\leq \frac{1}{2^{id}} \sum_{j\in[s_i],|Q_{i,j}|>0} \sum_{q\in Q_{i,j}} \frac{1}{|Q_i \cap \bar{B}_\infty(q,l_i)|} \leq \frac{1}{2^{id}} \sum_{j\in[s_i],|Q_{i,j}|>0} \sum_{q\in Q_{i,j}} \frac{1}{|Q_{i,j}|} \\
&= \frac{1}{2^{id}} \sum_{j\in[s_i],|Q_{i,j}|>0} \frac{|Q_{i,j}|}{|Q_{i,j}|} \leq \frac{s_i}{2^{id}} = \frac{2^{(i+1)d}}{2^{id}} = 2^d.
\end{aligned}$$

Then using the definitions of $\widetilde{C}_Q$ and r_q we have

$$\begin{aligned}
\widetilde{C}_Q &= \sum_{q\in Q} \max_{0<r\leq L} \frac{r^d}{L^d} \frac{1}{|Q \cap B_\infty(q,r)|} = \sum_{q\in Q} \frac{r_q^d}{L^d} \frac{1}{|Q \cap B_\infty(q,r_q)|} = \sum_{i\in A}\sum_{q\in Q_i} \frac{r_q^d}{L^d} \frac{1}{|Q \cap B_\infty(q,r)|} \\
&\leq \sum_{i\in A}\sum_{q\in Q_i} \frac{r_q^d}{L^d} \frac{1}{|Q_i \cap B_\infty(q,r)|} \leq \sum_{i\in A} 2^d = 2^d|A|.
\end{aligned}$$

We assert $r_q \geq Ln^{-\frac{1}{d}}$ for all $q \in Q$. This is because for any $r \in (0, Ln^{-\frac{1}{d}})$ we have

$$\frac{r^d}{L^d} \frac{n}{|Q \cap B_\infty(q,r)|} \leq \frac{L^d}{nL^d} \frac{n}{1} = 1 \leq \frac{L^d}{L^d} \frac{n}{|Q \cap B_\infty(q,L)|},$$

which implies the optimal $r_q \in [Ln^{-\frac{1}{d}}, L]$. Moreover, since $r_q \geq \min_{q'\in Q,\ q'\neq q} \|q - q'\|_\infty \geq \eta$, we have $r_q \geq \max(Ln^{-\frac{1}{d}}, \eta)$ for all $q \in Q$. If $i > \min\big(\log_2 \frac{L}{\eta}, \frac{1}{d}\log_2 n\big)$, then $\frac{L}{2^i} < \max(Ln^{-\frac{1}{d}}, \eta) \leq r_q$, and from the definition of Q_i and A we know $i \notin A$, which implies $|A| \leq 1 + \min\big(\log_2 \frac{L}{\eta}, \frac{1}{d}\log_2 n\big)$. Hence we obtain $\widetilde{C}_Q \leq 2^{d+1} \min\big(\log_2 \frac{L}{\eta}, \frac{1}{d}\log_2 n\big)$ and using $C_Q = (\widetilde{C}_Q)^{\frac{2}{2+d}}$ we prove the lemma. ◀

4 Strong Coresets for the Distance Between PL Curves

In this section, we study the distance $\mathtt{d}_Q$ defined on a subset of $\mathcal{S}(L)$: the collection of k-piecewise linear curves, and use the framework in [5] to construct a strong approximation for Q. This requires a bound on the shattering dimension, not possible for unrestricted

objects as in Section 3. We assume the multiset Q contains m distinct points $q_1, \cdots, q_m$, where each point q_i appears m_i times and $\sum_{i=1}^{m} m_i = n$. So, in this section Q will be viewed as a set $\{q_1, \cdots, q_m\}$ (not a multiset) and each point $q \in Q$ has a weight $w(q_i) = \frac{m_i}{n}$.

Suppose $\mathcal{T}_k := \{\gamma = \langle c_0, \cdots, c_k\rangle \mid c_i \in \mathbb{R}^d\}$ is the collection of all piecewise-linear curves with k line segments in $\mathbb{R}^d$. For $\gamma = \langle c_0, \cdots, c_k\rangle \in \mathcal{T}_k$, $\langle c_0, \cdots, c_k\rangle$ is the sequence of $k+1$ critical points of γ. The value $\mathsf{dist}(q, \gamma) = \inf_{p\in\gamma} \|p-q\|$, and function $f_{\gamma_1,\gamma_2}(q) = (\mathsf{dist}(q,\gamma_1) - \mathsf{dist}(q,\gamma_2))^2$ are defined as before. We now use weights $w(q_i) = \frac{m_i}{n}$ $\left(\sum_{q\in Q} w(q) = 1\right)$ and the resulting distance is $\mathtt{d}_Q(\gamma_1,\gamma_2) = \left(\sum_{q\in Q} w(q) f_{\gamma_1,\gamma_2}(q)\right)^{\frac{1}{2}}$.

For $L > \rho > 0$, $Q = \{q_1, \cdots, q_m\} \subset \mathbb{R}^d$, we define

$$\mathcal{X}_k^d(L,\rho) := \{(\gamma_1,\gamma_2) \in \mathcal{T}_k \times \mathcal{T}_k \mid \gamma_1,\gamma_2 \in \mathcal{S}(L),\ \mathtt{d}_Q(\gamma_1,\gamma_2) \geq \rho\}\,.$$

We next consider the sensitivity adjusted weights $w'(q) = \frac{\sigma(q)}{\mathfrak{S}_Q} w(q)$ and cost function $g_{\gamma_1,\gamma_2}(q) = \frac{1}{\sigma(q)} \frac{f_{\gamma_1,\gamma_2}(q)}{\bar{f}_{\gamma_1,\gamma_2}}$. These use the general bounds for sensitivity in Lemma 7 and Theorem 4, with as usual $\bar{f}_{\gamma_1,\gamma_2} = \sum_{q\in Q} w(q) f_{\gamma_1,\gamma_2}(q)$. These induce an adjusted range space $(Q, \mathcal{T}'_{k,d})$ where each element is defined

$$T_{\gamma_1,\gamma_2,\eta} = \{q \in Q \mid w'(q) g_{\gamma_1,\gamma_2}(q) \leq \eta,\ \gamma_1,\gamma_2 \in \mathcal{X}_k^d(L,\rho)\}.$$

Now to apply the strong coreset construction of Braverman *et al.* [5][Theorem 5.5] we only need to bound the shattering dimension of $(Q, \mathcal{T}'_{k,d})$.

Two recent results provide bounds on the VC dimension of range spaces related to trajectories. Given a range space $(X, \mathcal{R})$ with VC dimension ν and shattering dimension $\mathfrak{s}$, it is known that $\mathfrak{s} = O(\nu \log \nu)$ and $\nu = O(\mathfrak{s})$. So up to logarithmic factors these terms are bounded by each other. First Driemel *et al.* [12] shows VC dimension for a ground set of curves $\mathbb{X}_m$ of length m, with respect to metric balls around curves of length k, for various distance between curves. The most relevant case is where $m = 1$ (so the ground set are points like Q), and the Hausdorff distance is considered, where the VC dimension is bounded $O(d^2k^2 \log(km)) = O(k^2 \log k)$ for $d = 2$, and is at least $\Omega(\max\{k, \log m\}) = \Omega(k)$. Second, Matheny *et al.* [21] considered ground sets $\mathbb{X}_k$ of trajectories of length k, and ranges defined by geometric shapes which may intersect those trajectories anywhere to include them in a subset, but this result is also not directly relevant. Neither of these cases directly imply the results for our intended range space, since ours involves a pair of trajectories.

▶ **Lemma 9.** *The shattering dimension of range space $(Q, \mathcal{T}'_{k,d})$ is $O(k^3)$, for constant d.*

Proof. Suppose $(\gamma_1,\gamma_2) \in \mathcal{X}_k^d(L,\rho)$ and $\eta \geq 0$, where $\gamma_1 = \langle c_{1,0}, \cdots, c_{1,k}\rangle$ and $\gamma_2 = \langle c_{2,0}, \ldots, c_{2,k}\rangle$, then we can define the range $T_{\gamma_1,\gamma_2,\eta}$ as

$$\begin{aligned} T_{\gamma_1,\gamma_2,\eta} &:= \{q \in Q \mid w'(q) g_{\gamma_1,\gamma_2}(q) \leq \eta\} \\ &= \{q \in Q \mid w(q) f_{\gamma_1,\gamma_2}(q) \leq \mathfrak{S}_Q \bar{f}_{\gamma_1,\gamma_2} \eta\} \\ &= \{q \in Q \mid w(q)(\mathsf{dist}(q,\gamma_1) - \mathsf{dist}(q,\gamma_2))^2 \leq \mathfrak{S}_Q \bar{f}_{\gamma_1,\gamma_2} \eta\}. \end{aligned}$$

For a trajectory γ defined by critical points $c_0, c_1, \ldots, c_k$ for $j \in [k]$ define s_j as the segment between c_{j-1}, c_j and ℓ_j as the line extension of that segment. The distance between q and a segment s_j is illustrated in Figure 3 and defined

$$\xi_j := \mathsf{dist}(q, s_j) = \begin{cases} \mathsf{dist}(q, c_{j-1}), & \text{if } \langle c_j - c_{j-1},\ q - c_{j-1}\rangle \leq 0 \\ \mathsf{dist}(q, c_j), & \text{if } \langle c_{j-1} - c_j,\ q - c_j\rangle \leq 0 \\ \mathsf{dist}(q, \ell_j), & \text{otherwise} \end{cases}.$$

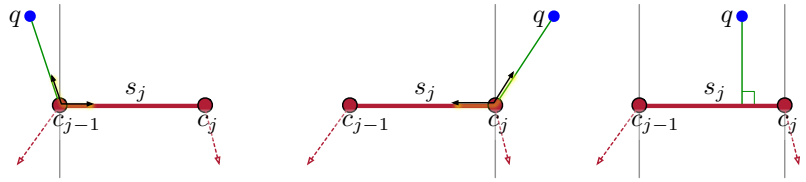


Figure 3 Illustration of the $\mathsf{dist}(q, s_j)$ from point q to segment s_j.

Then $\mathsf{dist}(q,\gamma) = \min_{j\in[k]} \xi_j$. For trajectories γ_1 and γ_2, specify these segment distances as $\xi_i^{(1)}$ and $\xi_i^{(2)}$, respectively. Then the expression for $T_{\gamma_1,\gamma_2,\eta}$ can be rewritten as

$$\begin{aligned}
T_{\gamma_1,\gamma_2,\eta} &= \{q \in Q \mid w'(q) g_{\gamma_1,\gamma_2}(q) \le \eta\} \\
&= \{q \in Q \mid w(q)(\min_{j\in[k]} \xi_j^{(1)} - \min_{j\in[k]} \xi_j^{(2)})^2 \le \mathfrak{S}_Q \bar{f}_{\gamma_1,\gamma_2}\eta\} \\
&= \cup_{j_1,j_2\in[k]} \{q \in Q \mid \xi_{j_1}^{(1)} \le \xi_j^{(1)}, \xi_{j_2}^{(2)} \le \xi_j^{(2)} \forall j \in [k], w(q)(\xi_{j_1}^{(1)} - \xi_{j_2}^{(2)})^2 \le \mathfrak{S}_Q \bar{f}_{\gamma_1,\gamma_2}\eta\} \\
&= \bigcup_{j_1,j_2\in[k]} \left(\begin{array}{l} \left(\cap_{j\in[k],j\neq j_1} \{q \in Q \mid \xi_{j_1}^{(1)} \le \xi_j^{(1)}\}\right) \\ \cap \left(\cap_{j\in[k],j\neq j_2} \{q \in Q \mid \xi_{j_2}^{(2)} \le \xi_j^{(2)}\}\right) \\ \cap \{q \in Q \mid \sqrt{w(q)}(\xi_{j_1}^{(1)} - \xi_{j_2}^{(2)}) \le (\mathfrak{S}_Q \bar{f}_{\gamma_1,\gamma_2}\eta)^{\frac{1}{2}}\} \\ \cap \{q \in Q \mid \sqrt{w(q)}(\xi_{j_2}^{(2)} - \xi_{j_1}^{(1)}) \le (\mathfrak{S}_Q \bar{f}_{\gamma_1,\gamma_2}\eta)^{\frac{1}{2}}\} \end{array} \right).
\end{aligned}$$

This means set $T_{\gamma_1,\gamma_2,\eta}$ can be decomposed as the union and intersection of $O(k^3)$ simply-defined subsets of Q. Specifically looking at the last line, this can be seen as the union over $O(k^2)$ sets (the outer union), and the first two lines are the intersection of $O(k)$ sets, and the last two lines inside the union are the intersection with one set each.

Next we argue that each of these $O(k^3)$ simply defined subsets of Q can be characterized as an element of a range space. By standard combinatorics [19, 3], the bound of the shattering dimension of the entire range space is $O(k^3)$ times the shattering dimension of any of these simple ranges spaces.

To get this simple range space shattering dimension bound, we can use a similar linearization method (see full version). For any simple range space $\mathcal{R}$ determined by the set decomposition of $T_{\gamma_1,\gamma_2,\eta}$, we can introduce new variables $c_0 \in \mathbb{R}, z, c \in \mathbb{R}^{d'}$, where z depends only on q, and c_0, c_i depend only on γ_1, γ_2 and r, and d' only depends on d. Here, Q is a fixed set and thus $\mathfrak{S}_Q$ is a constant. By introducing new variables we can construct an injective map $\varphi : Q \mapsto \mathbb{R}^{d'}$, s.t. $\varphi(q) = z$. There is also an injective map from $\mathcal{R}$ to $\{\{z \in \varphi(Q) \mid c_0 + z^T c \le 0\} \mid c_0 \in \mathbb{R}, c \in \mathbb{R}^{d'}\}$. Since the shattering dimension of the range space $(\mathbb{R}^{d'}, \mathcal{H}^{d'})$, where $\mathcal{H}^{d'} = \{h \text{ is a halfspace in } \mathbb{R}^{d'}\}$, is $O(d')$, we have the shattering dimension of$(Q, \mathcal{R})$ is $O(d') \le C_d$ where C_d is a positive constant depending only on d. Piecing this all together we obtain $C_d k^3$ bound for the shattering dimension of $(Q, \mathcal{T}'_{k,d})$. ◀

Now, we invoke Lemma 9 and [5][Theorem 5.5] to get a $(\rho, \varepsilon, \delta)$-strong coreset for $\mathcal{X}_k^d(L, \rho)$.

▶ **Theorem 10.** *Let $L > \rho > 0$, $Q \subset [0, L]^d$, and consider trajectory pairs $\mathcal{X}_k^d(L, \rho)$. Suppose $\sigma(q)$ and $\mathfrak{S}_Q$ are defined in Lemma 7 and Theorem 4 respectively. Then for $\delta, \varepsilon \in (0, 1)$ a σ-sensitive sampling of size $N = O(\frac{\mathfrak{S}_Q}{\varepsilon^2}(k^3 \log \mathfrak{S} + \log \frac{1}{\delta}))$ provides $\tilde{Q}$, a strong $(\rho, \varepsilon, \delta)$-coreset; that is with probability at least $1 - \delta$, for all pairs $\gamma_1, \gamma_2 \in \mathcal{X}_k^d(L, \rho)$ we have*

$$(1 - \varepsilon)\mathtt{d}_Q(\gamma_1, \gamma_2) \le \mathtt{d}_{\tilde{Q},W}(\gamma_1, \gamma_2) \le (1 + \varepsilon)\mathtt{d}_Q(\gamma_1, \gamma_2).$$

5 Trajectory Reconstruction

Let $\mathcal{T} := \{\gamma = \langle c_0, \cdots, c_k\rangle \mid c_i \in \mathbb{R}^2, k \geq 1\}$ be the set of all piecewise-linear curves in $\mathbb{R}^2$. Each curve in $\mathcal{T}$ is specified by a series of critical points $\langle c_0, c_1, \ldots, c_k\rangle$, and k line segments $s_1, s_2, \ldots, s_k$, where s_i is the line segment $\overline{c_{i-1}c_i}$. In this section we study how to recover γ from Q and $v_Q(\gamma)$ for $\gamma \in \mathcal{T}$.

Restrictions on curves and Q. For $\tau > 0$ we define a family of curves $\mathcal{T}_\tau \subset \mathcal{T}$ s.t. each $\gamma \in \mathcal{T}_\tau$ has two restrictions:

- (R1): Angles $\angle_{[c_{i-1},c_i,c_{i+1}]}$ at an internal critical point c_i is non-zero (i.e., in $(0, \pi)$).
- (R2): Each critical point c_i is τ-*separated*, that is the disk $B(c_i, \tau) = \{x \in \mathbb{R}^2 \mid \|x - c_i\| \leq \tau\}$ only intersects the two adjacent segments s_{i-1} and s_i of γ, or one adjacent segment for endpoints (i.e., only the s_1 for c_0 and s_k for c_k).

We next restrict that all curves (and Q) lie in region $\Omega \subset \mathbb{R}^2$. Let $\mathcal{T}_\tau(\Omega)$ be the subset of $\mathcal{T}_\tau$ where all curves γ have all critical points within Ω, and in particular, no $c_i \in \gamma$ is within a distance τ of the boundary of Ω.

To define Q, for $\eta > 0$, define an infinite grid $G_\eta = \{g_v \in \mathbb{R}^2 \mid g_\eta = \eta v \text{ for } v = (v_1, v_2) \in \mathbb{Z}^2\}$, where $\mathbb{Z}$ is all integers. Suppose $\eta \leq \frac{\tau}{32}$, then $Q = G_\eta \cap \Omega = \{q_1, \cdots, q_n\}$, $\gamma \in \mathcal{T}_\tau(\Omega)$, $v_i = \min_{p\in\gamma} \|q_i - p\|$ and $v_Q(\gamma) = (v_1, \ldots, v_n)$. We define some notations that are used in this section for the implied circle $C_i := \{x \in \mathbb{R}^2 | \ \|x - q_i\| = v_i\}$, the closed disk $B_i := \{x \in \mathbb{R}^2 | \ \|x - q_i\| \leq v_i\}$, and the open disk $\dot{B}_i := \{x \in \mathbb{R}^2 | \ \|x - q_i\| < v_i\}$ around each q_i or radius v_i. When the radius is specified as r (with perhaps $r \neq v_i$), then we, as follows, denote the associated circle $C_{i,r}$, closed disk $B_{i,r}$, and open disk $\dot{B}_{i,r}$ around q_i.

For Q, $\gamma \in \mathcal{T}_\tau(\Omega)$ and $v_Q(\gamma)$ we have the following three observations.

- (O1): In any disk with radius less than τ, there is at most one critical point of γ; by (R2).
- (O2): If a point moves along γ, it can only stop or change direction at critical points of γ.
- (O3): For any $q_i \in Q$, γ cannot go into $\dot{B}_i$. Moreover, C_i must contain at least one point of γ, and if this point is not a critical point, then γ must be tangent to C_i at this point.

The restriction (R2) only implies if there is a critical point of γ, then in its neighborhood γ has at most two line segments. However, if there is no critical point in a region, then the shape of γ can be very complicated in this region, so we need to first identify the regions that contain a critical point.

These observations form the basis for the algorithm and its proof of correctness. We next describe the algorithm, state the main results, and provide intuition for the proofs. For space, some pseudocode and full proofs which rely heavily on case analysis are in the full version.

Recovery algorithm. The entire algorithm is overviewed in Algorithm 1. For each critical point $c \in \gamma$, there exists $q \in Q$ such that $\mathsf{dist}(q, c) < \eta$. So to recover γ, we first traverse $\{q_i \in Q \mid v_i < \eta\}$ and use $\textsc{isCritical}(q_i)$ to solve the decision problem of if there is a critical point in $B_{i,3\eta}$. Whenever there is a critical point in $B_{i,3\eta}$, we then use $\textsc{FindCritical}(q_i)$ to find it – collectively, this finds all critical points of γ. Finally, we use $\textsc{DetermineOrder}$ (Algorithm 2) to determine the order of all critical points of γ, which recovers γ.

Existence of critical points. In $\textsc{isCritical}(q_i)$ we consider the common tangent line of C_i and C_j for all q_j in a neighborhood of q_i. If no common tangent line can go through $B_{i,3\eta}$ without going into the interior of any other circle centered in $B_{i,3\eta}$, unlike Figure 4(Left), then it implies there is a critical point of γ in $B_{i,3\eta}$. This can be confirmed checking the possible tangent lines for circles centered at grid points in $B_{i,3\eta}$ and around q_i; cases for endpoints and internal points are illustrated in Figure 4(Center,Right).

Algorithm 1 Recover $\gamma \in \mathcal{T}_\tau(L)$ from Q and $v_Q(\gamma)$.

Initialize $Q_\eta := \{q_i \in Q \mid v_i < \eta\}$, close set $Q_{\mathrm{r}} := \emptyset$, endpoints $E = \emptyset$ and critical points $A := \emptyset$.
for each $q_i \in Q_\eta$ **do**
  if $q_i \in Q_{\mathrm{r}}$ or IsCritical(q_i)=False **then**
    continue
  Let $(c, S) :=$ FindCritical(q_i).
  if $|S| = 1$ **then**
    $E := E \cup \{(c, S)\}$. // c is an endpoint of γ
  Let $A := A \cup \{(c, S)\}$ and $Q_{\mathrm{r}} := Q_{\mathrm{r}} \cup \big(Q_\eta \cap B_{c,16\eta}\big)$. // aggregate critical points
return $\gamma :=$ DetermineOrder(E, A)

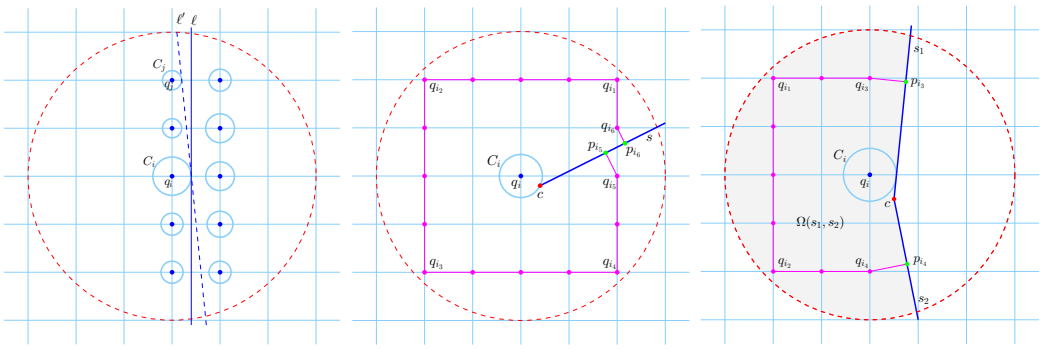


Figure 4 Determining no critical point (Left), endpoint (Center), or internal critical point (Right).

▶ **Lemma 11.** *Suppose $q_i \in Q$ and $v_i < \eta$. If* IsCritical(q_i) *returns* True*, then there must be a critical point of γ in $B_{i,3\eta}$. Moreover, for any critical point $c \in \gamma$ there exists some $q_i \in Q$ such that $v_i < \eta$ and* IsCritical(q_i) *returns* True *for the input q_i.*

Finding a critical point. If there is a critical point c in $B_{i,3\eta}$, then using (R2) we know in the neighborhood of c, γ has a particular pattern: it either has one line segment, or two line segments. We will need two straightforward subfunctions:

- FCT (*Find Common Tangents*) takes in three grid points q_i, q_j, q_k, and returns the all common tangent lines of C_j and C_k which do not intersect the interior of disks $\dot{B}_l$ of a disk associated with a point $q_l \in Q_{i,8\eta}$. This generates a feasible superset of possible nearby line segments which may be part of γ.
- MOS (*Merge-Overlapping-Segments*) takes a set of line segments, and returns a smaller set, merging overlapping segments. This combines the just generated potential line segments of γ.

Now in FindCritical(q_i) for each pair $q_j, q_k \in B_{i,8\eta}$, we first use FCS to find the common tangent line of C_j, C_k that could be segments of γ, and then use MOS to reduce this set down to a minimal set of possibilities S_m. By definition, there must be a critical point c, and thus can be at most 2 actual segments of γ within $B_{i,8\eta}$, so we can then refine S_m. We first check if c is an endpoint, in which case there must be only one valid segment. If not, then there must be 2, and we need to consider all pairs in S_m. This check can be done by verifying that *every* C_k for $q_k \in Q_{i,8\eta}$ is tangent to the associated ray $\mathsf{ray}(s)$ (for an endpoint) or for the associated rays $\mathsf{ray}(s)$ and $\mathsf{ray}(s')$ for their associated segment pairs (for an internal critical point). Some of these cases are illustrated in Figure 5.

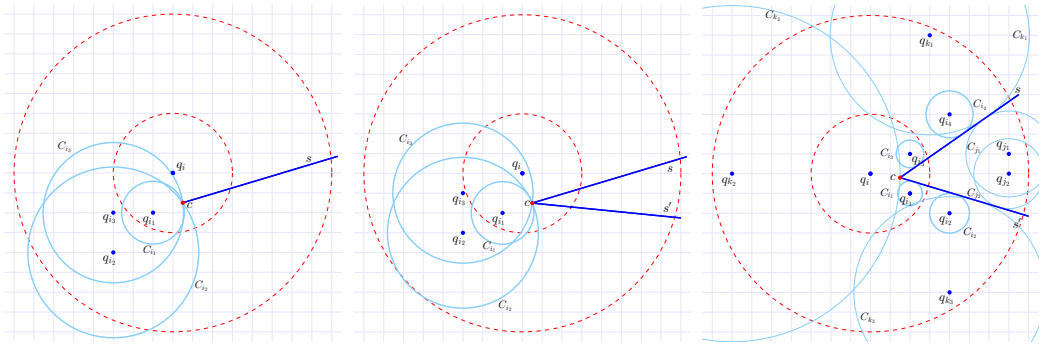


Figure 5 Illustration of how $Q \cap B_{i,8\eta}$ is sufficient to recover a critical point c in $B_{i,3\eta}$ for the c and endpoint (Left), or an internal point with small angle (Center) or large angle (Right).

▶ **Lemma 12.** *Suppose $c' \in B_{i,3\eta}$ is a critical point of γ, and (c, S) is the output of* FindCritical(q_i)*, then $c = c'$. Moreover, $|S| = 1$ if and only if c is an endpoint of γ.*

Using isCritical and FindCritical we can find all critical points (E, A) with associated line segments of γ, so the final step is to use function DetermineOrder(E, A) (Algorithm 2) to determine their order, as we argue it will completely recover γ.

Algorithm 2 DetermineOrder(E, A): Determine the order of critical points.

Choose any $(c_0, S_0) \in E$, let $k = |A| - 1$, $A := A \setminus \{(c_0, S_0)\}$, $s_1 \in S_0$ and $\gamma := \langle c_0 \rangle$.
for $i = 1$ **to** k **do**
  Find closest c from $(c, S) \in A$ to c_{i-1} so c is on $\mathsf{ray}(s_i)$, and let $A := A \setminus \{(c, S)\}$.
  Append $c_i = c$ to γ, and if $i < k$ then let $s_{i+1} = s$ where $s \in S$ is not parallel with s_i.
return γ

▶ **Theorem 13.** *Suppose $Q = G_\eta \cap \Omega$, $\eta \leq \frac{\tau}{32}$, and $v_Q(\gamma)$ is generated by Q and $\gamma \in \mathcal{T}_\tau(\Omega)$ with k segments, then Algorithm 1 can recover γ from $v_Q(\gamma)$ in $O(|Q| + k^2)$ time.*

References

1 Ann Arbor Algorithms. K-graph. Technical report, `https://github.com/aaalgo/kgraph`, 2018.
2 Nina Amenta, Sunghee Choi, and Ravi Krishna Kolluri. The power crust. In *Proceedings of the sixth ACM symposium on Solid modeling and applications*, 2001.
3 Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.
4 Christos Boutsidis, Michael W Mahoney, and Petros Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2009.
5 Vladimir Braverman, Dan Feldman, and Harry Lang. New frameworks for offline and streaming coreset constructions, 2016. `arXiv:1612.00889`.
6 Frederic Chazal and David Cohen-Steiner. Geometric inference. URL: `https://geometrica.saclay.inria.fr/team/Fred.Chazal/papers/GeomInference5.pdf`.
7 Frederic Chazal, David Cohen-Steiner, and Andre Lieutier. A sampling theory for compact sets in euclidean space. *DCG*, 41:461–479, 2009.
8 Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. Geometric inference for probability measures. *Foundations of Computational Mathematics*, pages 1–19, 2010.

9 Frédéric Chazal and Andre Lieutier. The "λ-medial axis". *Graphical Models*, 67:304–331, 2005.

10 Michael B. Cohen, Cameron Musco, and Christopher Musco. Input sparsity time low-rank approximation via ridge leverage score sampling. In *ACM-SIAM Symposium on Discrete Algorithms*, 2017.

11 Michael B. Cohen, Cameron Musco, and Jakub Pachocki. Online row sampling. In *International Workshop on Approximation, Randomization, and Combinatorial Optimization*, 2016.

12 Anne Driemel, Jeff M. Phillips, and Ioannis Psarros. On the vc dimension of metric balls under frechet and hausdorff distances. In *International Symposium on Computational Geometry*, 2019.

13 Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of statistical leverage. *Journal of Machine Learning Research*, 13:3475–3506, 2012.

14 Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM Journal of MAtrix Analysis and Applications*, 30:844–881, 2008.

15 Herbert Edelsbrunner and Ernst P. Mücke. Three-dimensional alpha shapes. *ACM Transactions on Graphics*, 13:43–72, 1994.

16 Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings ACM Symposium on Theory of Computing*, 2011.

17 Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, PCA, and projective clustering. In *Proceedings 24th ACM-SIAM Symposium on Discrete Algorithms*, 2013.

18 Dan Feldman and Lenard J. Schulman. Data reduction for weighted and outlier-resistant clustering. In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, 2012.

19 S. Har-Peled. *Geometric Approximation Algorithms*. Mathematical Surveys and Monographs. American Mathematical Society, 2011.

20 Michael Langberg and Leonard J. Schulman. Universal ε-approximators for integrals. In *SODA*, pages 598–607, 2010.

21 Michael Matheny, Dong Xie, and Jeff M. Phillips. Scalable spatial scan statistics for trajectories, 2019. `arXiv:1906.01693`.

22 Cameron Musco and Christopher Musco. Recursive sampling for the Nyström method. In *NIPS*, 2017.

23 Jeff M. Phillips and Pingfan Tang. Simple distances for trajectories via landmarks. In *SIGSPATIAL*. (long version: `arXiv:1804.11284`), 2019.

24 Jeff M. Phillips, Bei Wang, and Yan Zheng. Geomtric inference on kernel density estimates. In *SOCG*, 2015.

25 Ilya Razenshteyn and Ludwig Schmidt. Falconn-fast lookups of cosine and other nearest neighbors. `https://falconn-lib.org`, 2018.

26 Kasturi Varadarajan and Xin Xiao. On the sensitivity of shape fitting problems. In Deepak D'Souza, Telikepalli Kavitha, and Jaikumar Radhakrishnan, editors, *IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2012)*, pages 486–497, Dagstuhl, Germany, 2012. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. `doi:10.4230/LIPIcs.FSTTCS.2012.486`.