Geometric Summaries: Coresets (and Beyond)

Pankaj K. Agarwal

Duke University

Large Data Sets

S: Set of n *points* in $\mathbb{R}^d$

- Both n and d are becoming large
- Other geometric objects

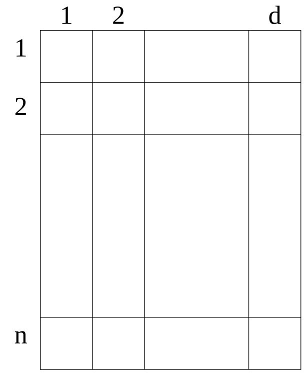
☆ Intractability

- NP-, PSPACE-hardness
- Even quadratic-time algorithms impractical
- Curse of dimensionality: exponential dependency on d

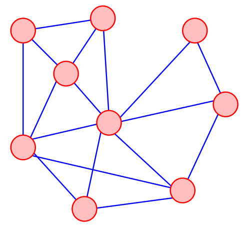
☆ Approximation algorithms

- *Work with a sparse representation (summary) of S*

Summaries



- ☆ Sampling, *coresets*
  - Choose a small subset K of S
  - Choosing a subset of rows
- ☆ Dimension reduction
  - Choosing a subset of columns
  - Multiply by a $d \times k$ matrix
- ☆ Sparsification of $S \times S$
  - Similarity matching, classification
  - Spanners
  - Bipartite clique cover
  - WSPD, SSPD



Overview

- ☆ **Part I: Early Results**
  - Coresets, ε-kernels
- ☆ **Part II: Recent Results**
  - Dynamic coresets
  - Coresets in streaming model
- ☆ **Part III: Other Summaries**
  - Coresets for nonextent measures
  - Spanners

Random Sampling

[Vapnik-Chervonenkis]

- ☆ $X = (S, R)$, $R \subseteq 2^S$: Set system (range space)
  - δ: VC-dimension of X
- ☆ $A \subseteq S$ *ε-approximation* if for all $r \in R$

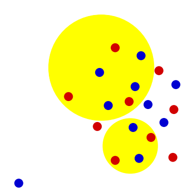
$$\left| \frac{|r|}{|S|} - \frac{|r \cap A|}{|A|} \right| \leq \varepsilon$$

- ☆ A random subset $A \subset S$ of size $\frac{\delta^2}{\varepsilon^2} \log \frac{\delta}{\varepsilon}$ is an ε-approximation of S with high probability
- ☆ Efficient deterministic algorithms for computing an ε-approximation [Matoušek, Chazelle]

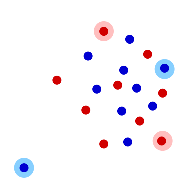
ε-Approximations

A: ε-approximation of S

- ☆ A is a coreset of S in a *combinatorial/statistical* sense
  - E.g. Approximate range counting
  - Approximates the distribution



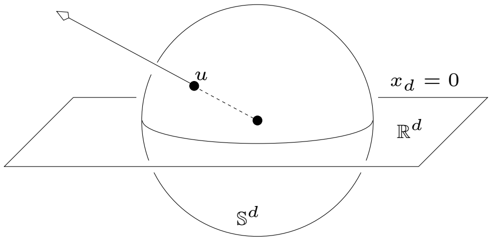
- ☆ A is *not* a coreset of S in a metric/geometric sense
  - $\text{diam}(A)$ does not approximate $\text{diam}(S)$
  - A best-fit circle for A does not approximate the best-fit circle for S

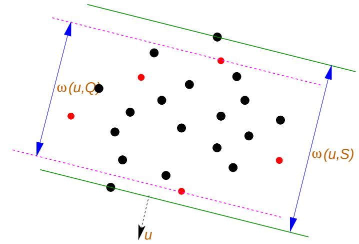


What about other sampling schemes?

ε-Kernels

S: Set of points in $\mathbb{R}^d$





Directional width: For $u \in \mathbb{S}^{d-1}$,

$$\omega(u, S) = \max_{p \in S} \langle u, p \rangle - \min_{p \in S} \langle u, p \rangle$$

ε-kernel: $Q \subseteq S$ is an ε-kernel of S if

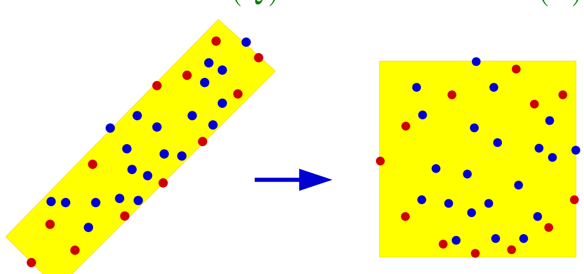
$$\omega(u, Q) \geq (1 - \varepsilon)\omega(u, S) \qquad \forall u \in \mathbb{S}^{d-1}$$

Computing ε-Kernels

Theorem A: [AHV, Ch, YAPV] $S \subseteq \mathbb{R}^d$, $\varepsilon > 0$. *An* ε*-kernel of* S *of size* $1/\varepsilon^{(d-1)/2}$ *can be computed in time* $n + 1/\varepsilon^{d-3/2}$.

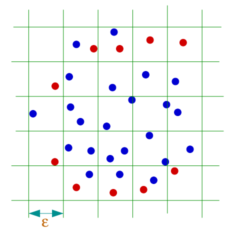
Lemma 1: $\exists$ affine transform M s.t.

- ☆ Unit hypercube $[-1, +1]^d$ is the smallest box enclosing S
- ☆ $M(S)$ is fat
- ☆ Q is an ε-kernel of $S \Leftrightarrow M(Q)$ is an ε-kernel of $M(S)$



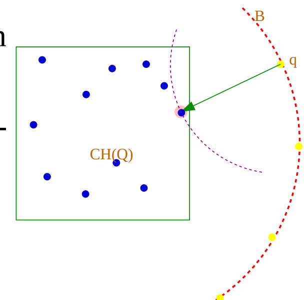
Computing ε-Kernels

Lemma 2: S: Set of n fat points in $[-1,+1]^d$, $\varepsilon > 0$. An ε-kernel of S of size $1/\varepsilon^{(d-1)/2}$ can be computed in time $n + 1/\varepsilon^{d-3/2}$.



Sketch: Algorithm in two phases

- ☆ Compute $1/\varepsilon^{d-1}$-size approximation Q
- ☆ Draw a sphere B of radius 2 centered at origin
- ☆ Draw a grid of size $1/\varepsilon^{(d-1)/2}$ on B
- ☆ For each grid point q, select its nearest neighbor in Q
  - Suffices to compute approximate NN
  - Use Arya-Mount ANN software library



Applications of ε-Kernels

$n + 1/\varepsilon^{O(1)}$-time approximation algorithms for computing

- ☆ extent measures: diameter, width
- ☆ smallest enclosing convex shapes
  - ball, ellipse
  - rectangle, simplex,

    ⋮

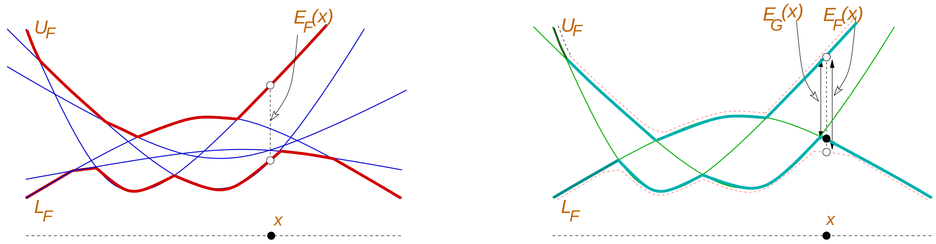
Fails to approximate

- ☆ Extent of moving points
- ☆ Smallest enclosing non-convex shapes
  - Minimum-width annulus
  - Minimum-width cylindrical shell

Extents of Functions

☆ $F = \{f_1, \dots, f_n\}$: d-variate functions

- U_F: Upper envelope of F $U_F(x) = \max_i f_i(x)$
- L_F: Lower envelope of F $L_F(x) = \min_i f_i(x)$



Extent of F:

$$E_F(x) = U_F(x) - L_F(x)$$

ε-kernel: $G \subseteq F$ is an ε-kernel of F if

$$(1 - \varepsilon)E_F(x) \le E_G(x) \qquad \forall x \in \mathbb{R}^d$$

ε-Kernels of Polynomials

$F = \{f_1, \dots, f_n\}$: d-variate polynomials

Linearization [Yao-Yao, A.-Matoušek, ...]

☆ Map $\varphi : \mathbb{R}^d \to \mathbb{R}^k$, $\varphi(x) = (\varphi_1(x), \dots, \varphi_k(x))$

☆ Each f_i maps to a k-variate linear function h_i;
$f_(x) > 0 \Leftrightarrow h_i(\varphi(x)) > 0$

☆ k: Dimension of linearization

Using linearization + duality:

Theorem C: *F: a family of n d-variate polynomials, k: dimension of linearization, $\varepsilon > 0$. We can compute an ε-kernel of F of size $1/\varepsilon^{k/2}$ in time $n + 1/\varepsilon^{k-1/2}$.*

Applications

Nonconvex-shape fitting

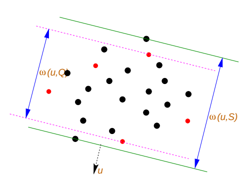
- ☆ $n + 1/\varepsilon^{O(1)}$ time approximation algorithms
  - minimum-width annulus
  - cylindrical shell
- ☆ Exact algorithms quite expensive

Kinetic data structures (KDS)

- ☆ maintaining approximate
  - diameter, width
  - smallest enclosing shape: box, ball, ellipse
  - # events: $1/\varepsilon^{O(1)}$, update time: $\log^{O(1)} 1/\varepsilon$
- ☆ Exact KDS require $\Omega(n^2)$ events

Robust Kernels

- ☆ Notion of ε-kernel is susceptible to outliers!
- ☆ $S^k[u]$: kth extremal point in direction u



$$\omega_{k,\ell}(u) = \langle u, S^k[u] \rangle - \langle u, S^k[-u] \rangle$$

(k, ε)**-kernel:** $Q \subseteq S$ is (k, ε)-kernel if

$$\omega_{a,b}(u, Q) \geq (1 - \varepsilon)\omega_{a,b}(u, S) \quad \forall u \in \mathbb{S}^{d-1}, a, b \leq k$$

$\delta = \varepsilon/4$, $S_0 = S$

for $0 \leq i \leq 2k$ do

T_i: δ-kernel of P_i; $\quad S_{i+1} = S_i \setminus T_i$

- ☆ $Q = \bigcup_{i=0}^{2k} T_i \; |Q| = k/\varepsilon^{(d-1)/2}$

Theorem G: *Q is a (k, ε)-kernel.*

Coresets in High Dimensions

- ☆ Computing $(d/\varepsilon)^{O(1)}$-size coresets in high dimensions [Bădoiu, Har-Peled, Indyk], [Bădoiu, Clarkson], [Har-Peled, Varadarajan], [Kumar, Mitchell, Yildirim], [Kumar, Yildirim]
  - Smallest enclosing ball $\lceil 1/\varepsilon \rceil$
  - Smallest enclosing ellipsoid $O(d/\varepsilon)$
  - 1-median $1/\varepsilon^{O(1)}$
- ☆ Relation to the Frank-Wolfe algorithm for quadratic programming [Clarkson]
- ☆ Coresets for distance between two polytopes [Gärtner, Jaggi]
- ☆ Computing coresets for clustering [Bădoiu, Har-Peled, Indyk], [Har-Peled, Ke], [Ke], [A., Procopiuc, Varadarajan]
  - k-centers, k-medians, k-line-centers

PART II

- ☆ Dynamic coresets: *insertion/deletion of points*
  - Linear size
  - Small update time
- ☆ Corsets in streaming model: *insertion only*
  - Small Size: independent of n
  - Small update time

Dynamic ε-Kernels

Maintain ε-kernel as points are inserted and deleted!

[A., Har-Peled, Varadarajan]
Size: $1/\varepsilon^{(d-1)/2}$, Update time: $(\log n/\varepsilon)^{O(d)}$

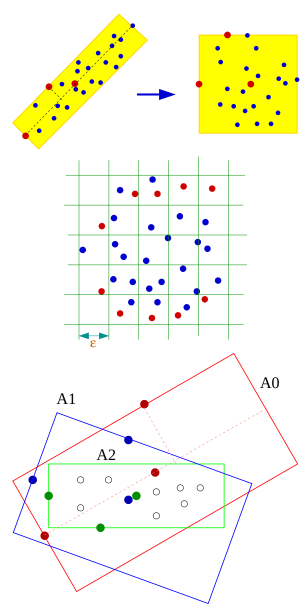
[Chan]
Size: $1/\varepsilon^{(d-1)/2}$, Update time: $(1/\varepsilon^{d-1}) \log n$

Algorithms work in two stages:

- ☆ Maintains a $(\varepsilon/2)$-kernel $\mathcal{L}$ of size $1/\varepsilon^{d-1}$
- ☆ Computes a $(\varepsilon/2)$-kernel $\mathcal{K}$ of $\mathcal{L}$
- ☆ $\mathcal{K}$: ε-kernel of S

Chan's $O(1/\varepsilon^{d-1})$-Size Algorithm

- ☆ *Anchor points*
  - $a_0, \ldots, a_d$
  - Define affine transform
  - Define bounding box B
- ☆ Anchor points fixed: Update is easy
- ☆ Updating anchors
  - Partition S into $\log n$ layers $S_0, S_1, \ldots, S_u$
  - $|S_i| \geq \alpha |S_{i-1}|$
  - $\bigcup_{j<i} S_j$ acts as anchors for S_i
  - $\bigcup_{j<i} S_j \neq \emptyset$: Use above algorithm



Stable Dynamic Algorithm

$\mathcal{K}$ may completely change after each update!

- ☆ [A., Phillips, Yu] Maintain an ε-kernel $\mathcal{K}$
  - Size: $1/\varepsilon^{(d-1)/2}$, Update time: $\log n + 1/\varepsilon^{(d-1)/2}$
  - $O(1)$ changes in $\mathcal{K}$ at each update
- ☆ Main idea
  - Fixed anchors: stable updates for the $1/\varepsilon^{(d-1)/2}$ size ε-kernel
  - Stable version of Chan's $1/\varepsilon^{d-1}$-size algorithm
  - "Gradual" morphing of the two algorithms

Stable (deterministic) algorithms for maintaining ε-nets and ε-approximations?

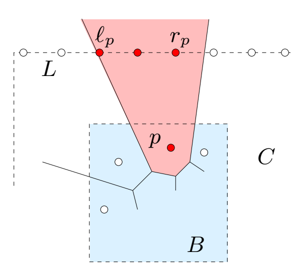
Streaming Model

- ☆ S: Stream of points in $\mathbb{R}^2$; points arrive one-by-one
- ☆ *Maintain the ε-kernel using $1/\varepsilon^{O(1)}$ space*

[A., Har-Peled, Varadarajan], [Chan], [A., Yu]

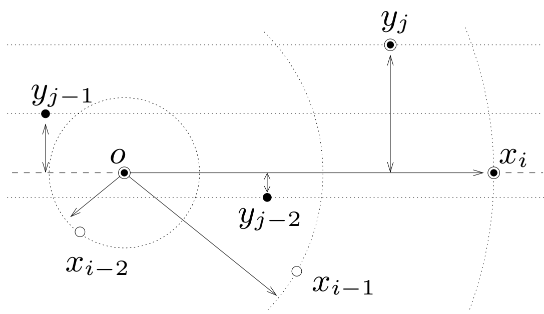
Theorem F [AY]: *ε-kernel in $\mathbb{R}^2$ can be in the streaming model using $O(1/\sqrt{\varepsilon})$ space and $O(\log(1/\varepsilon))$ update time.*

- ☆ Problem is easy as long as anchor points fixed
- ☆ Keep track of NN of each grid point
  - Update time: $\log(1/\varepsilon)$



Streaming: Algorithm Overview

Maintaining anchor points: epochs and subepochs



- ☆ o: first point in the stream
- ☆ x_i: first point in the ith epoch
- ☆ y_j: first point in the jth subepoch of the current epoch
- ☆ x_i starts a new epoch if $\|ox_i\| > 2\|ox_{i-1}\|$
- ☆ y_j starts a new subepoch if $d(y_j, \ell(o,a)) > 2d(y_{j-1}, \ell(o,a))$

Streaming: Algorithm Overview

- ☆ Maintain ε-kernels for $\log(1/\varepsilon)$ epochs
  - Maintain ε-kernels for $\log(1/\varepsilon)$ subepochs within each epoch
  - Points in earlier epochs are too close to o
  - Points in earlier subepochs are too close to the line ox_i
  - Size: $(1/\sqrt{\varepsilon})\log^2(1/\varepsilon)$
- ☆ Prune coresets from older epochs and subepochs
  - Size: $(1/\sqrt{\varepsilon})$

Streaming in High Dimensions

- ☆ d is large and part of the input
- ☆ [A. Raghvendra] For $\varepsilon = d^{1/3}$, size of ε-kernel is $\Omega(\exp(d^{1/3}))$.
- ☆ Coresets of size $(d/\varepsilon)^{O(1)}$ for some problems
- ☆ *Are there streaming algorithms that use $(d/\varepsilon)^{O(1)}$ space to maintain coresets?*

Streaming in High Dimensions

- ☆ **Minimum enclosing ball (MEB)** [Chan, Zarrabi-Zadeh]
  - Maintains a single ball
  - Size: $O(d)$
  - $(1+\sqrt{2})/2$-approximation
  - Bound is tight for any structure that maintains only one ball
- ☆ **Diameter** [Indyk]
  - c-approximation, for $c > \sqrt{2}$
  - Size: $dn^{1/(c^2-1)}$
  - Update time: $dn^{1/(c^2-1)}$

Streaming in High Dimensions

[A., Raghvendra]

☆ **Diameter**

- $(\sqrt{2} - 1/d^{1/3})$-approximation, size: $\Omega(\exp(d^{1/3}))$
- $(\sqrt{2} + \varepsilon)$-approximation, size: $O((d/\varepsilon^3)\log(1/\varepsilon))$

☆ **Minimum enclosing ball (MEB)**

- $((1+\sqrt{2})/2 - 1/d^{1/3})$-approximation, size: $\Omega(\exp(d^{1/3}))$
- $((1+\sqrt{3})/2 + \varepsilon)$-approximation, size: $O((d/\varepsilon^3)\log(1/\varepsilon))$

☆ Lower bounds are proved using *communication complexity*

☆ Upper bounds based on a notion of *blurred ball cover*

Lower Bound: Set-Disjointness Problem

☆ $U : [1 : k]$

☆ Alice has a set $A \subseteq U$

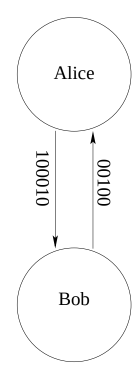
☆ Bob has a set $B \subseteq U$

☆ Alice & Bob communicate to determine

Is $A \cap B = \emptyset$?

☆ Communication Complexity: # bits communicated

☆ Communication complexity = $\Omega(k)$
[Kalyansundaram-Schnitger]



Lower Bound: Diameter

Lemma: $\exists K \subset \mathbb{S}^{d-1}$ s.t.
(i) $|K| = \exp(d^{1/3})$, (ii) $p \in K \Rightarrow -p \in K$,
(iii) $p, q \in K, p \neq q \Rightarrow \|pq\| \approx \sqrt{2}$

- ☆ $X = K \cap \{x_d \geq 0\}$, $\phi : [1:k] \to X$, $k \approx \exp(d^{1/3})$
- ☆ $\mathbb{D}$: Maintains $\sqrt{2}$-diameter in the streaming model
  - Returns $s, t \in S$ s.t. $\forall p, q \in S$ $\|pq\| \leq \sqrt{2}\|st\|$

Alice
- ☆ $\forall a \in A$ insert $\phi(a)$ to $\mathbb{D}$
- ☆ Communicate $\mathbb{D}$ to Bob

Bob
- ☆ $\forall b \in B$ insert $-\phi(b)$ to $\mathbb{D}$
- ☆ If $\mathbb{D}$ returns an antipodal pair
  Return $A \cap B \neq \emptyset$

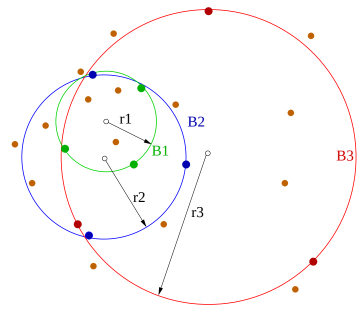
Communication complexity: Size($\mathbb{D}$)

Lower Bound: Diameter

- ☆ $\text{diam}(\phi(A)) \approx \sqrt{2}$
- ☆ $\text{diam}(\phi(A) \cup -\phi(B)) \approx \begin{cases} 2 & A \cap B \neq \emptyset \\ \sqrt{2} & A \cap B = \emptyset \end{cases}$
- ☆ $\mathbb{D}$ can distinguish the two case
- ☆ Communication complexity = Size($\mathbb{D}$) = $\Omega(k) = \Omega(\exp(d^{1/3}))$

Blurred Ball Cover

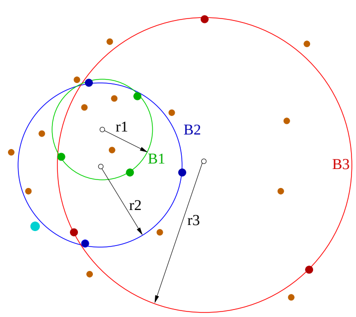
- ☆ S: set of points
- ☆ $\mathcal{K} = \{K_1, \ldots, K_u\}$, $K_i \subseteq S$, $|K_i| \approx 1/\varepsilon$
- ☆ $B_i = \text{MEB}(K_i)$, $r_i = r(B_i)$
- ☆ K: ε-blurred ball cover if
  - $r_{i+1} \geq (1+\varepsilon^2) r_i$
  - $\forall j \leq i$, $K_j \subseteq B_i$
  - $S \subset \bigcup_i (1+\varepsilon) B_i$



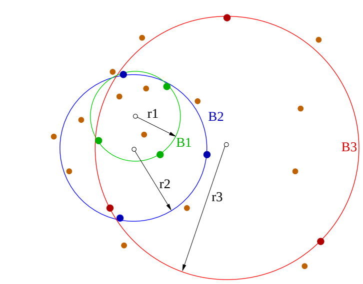
$$r_u \leq r_1/\varepsilon \;\Rightarrow\; u \approx \frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon}, \sum_i |K_i| \approx \frac{1}{\varepsilon^3}\log\frac{1}{\varepsilon}$$

Inserting a Point

Case I: $\exists i, p \in (1+\varepsilon) B_i$

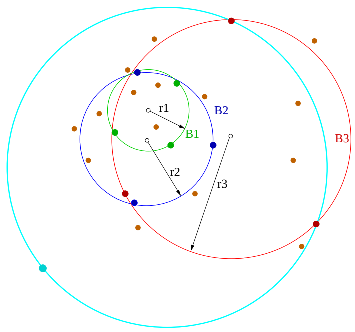


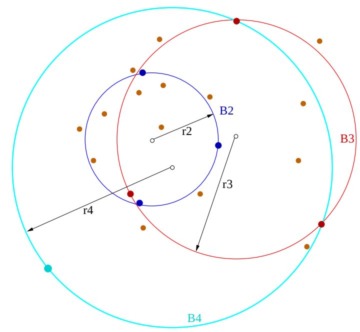
Case II: $\forall i, p \notin (1+\varepsilon) B_i$



Updating Blurred Ball Cover

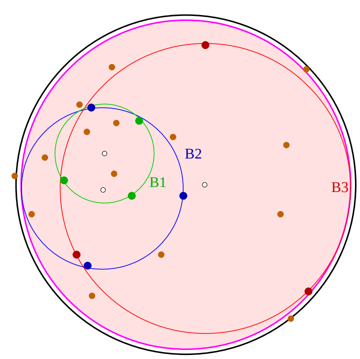
- $B^*, K^* = \text{APPROX\_MEB}(\bigcup \mathcal{K} \cup \{p\}, \varepsilon/2)$
- Insert K^* into $\mathcal{K}$, delete $\{K_i \mid r_i < \varepsilon r(B^*)\}$





Minimum Enclosing Ball

- Return $\mathcal{B} = \text{MEB}(B_1, \ldots, B_u)$
- $S \subset (1 + \varepsilon/2)\mathcal{B}$
- $r(\mathcal{B})/r(\text{MEB}(S)) \le \dfrac{1+\sqrt{3}}{2} + \varepsilon$



PART III: Coresets for Shortest Paths

- ☆ $\mathcal{P} = \{P_1, \ldots, P_k\}$: Pairwise-disjoint convex obstacles in $\mathbb{R}^3$
- ☆ $\mathcal{F}(\mathcal{P}) = \mathbb{R}^3 \setminus \bigcup P$: Free space
- ☆ For $s, t \in \mathcal{F}(\mathcal{P})$, $d_{\mathcal{P}}(s,t)$: length of the collision-free shortest path

Given $\varepsilon > 0$, is there a small-sketch $\mathcal{Q} = \{Q_1, \ldots, Q_k\}$,

(i) $P_i \subseteq Q_i$

(ii) $\sum_i |Q_i|$ *small*

(iii) $\forall s, t \in \mathcal{F}(\mathcal{Q})$ $d_{\mathcal{Q}}(s,t) \leq (1+\varepsilon) d_{\mathcal{P}}(s,t)$

DUKE COMPUTER SCIENCE

Coresets for Shortest Paths

[A., Raghvendra, Yu]

$d = 2$**:**

- ☆ $\sum_i |Q_i| = \Theta(k/\sqrt{\varepsilon})$

$d = 3$**:**

- ☆ No small-size sketch exists if neither s nor t is given
- ☆ If s is fixed (but t is arbitrary)
  - $\sum_i |Q_i| \approx O((k/\varepsilon)^3)$
  - $\sum_i |Q_i| \approx \Omega(k^2)$

Is there a binary spce partition of a set of disjoint k convex objects in $\mathbb{R}^3$ of size $O(k^2)$?

DUKE COMPUTER SCIENCE

Spanners in 2D

- ☆ $\mathcal{P}$: k pairwise disjoint polygons in $\mathbb{R}^2$
- ☆ n: # vertices in $\mathcal{P}$
- ☆ $S = \{x_1, \ldots, x_n\}$: n points in $\mathbb{R}^2$

Geodesic-distance graph: $\mathbb{G}(\mathcal{P}, S) = S \times S$,

$$w(x_i, x_j) = d_{\mathcal{P}}(x_i, x_j)$$

Visibility graph: $\mathbb{V}(\mathcal{P}, S) = (S, E)$,

$$(x_i, x_j) \in E \Leftrightarrow x_i x_j \subset \mathcal{F}(\mathcal{P}) \quad w(x_i, x_j) = |x_i - x_j|$$

Given a graph G, $H \subseteq G$ *t-spanner* of G if

$$d_H(u, v) \leq t \cdot d_G(u, v) \; \forall u, v \in V(G)$$

Are there small-size spanners of $\mathbb{G}$ and $\mathbb{V}$?

Spanners

[Abam, A., de Berg]: Work in progress

Visibility graph

- ☆ F is a simple polygon
  - $t \leq 3 - \varepsilon$, $|H| = \Omega(n^2)$
  - $t \geq 6 + \varepsilon$, $|H| \approx n^{4/3}$
- ☆ F has holes
  - $t \leq 5 - \varepsilon$, $|H| = \Omega(n^{4/3})$

Geodesic distance graph

- ☆ F is a simple polygon
  - $t \leq 2 - \varepsilon$, $|H| = \Omega(n^2)$
  - $t \geq 3 + \varepsilon$, $|H| \approx n \log^2 n$
- ☆ Some weak results when P has holes