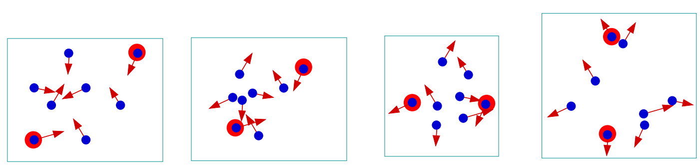
Geometric Approximation Using Core-Sets

Pankaj K. Agarwal

Department of Computer Science
Duke University

Kinetic Geometry



S: Set of n moving points in $\mathbb{R}^2$

- $p_i = a_i + b_i t$

Maintain the diameter (width, smallest enclosing disk) of S.

☆ [A., Guibas, Hershberger, Veach]

- Diametral pair can change $\Theta(n^2)$ times
- Kinetic data structure with $\approx n^2$ events

☆ *Can we maintain the approximate diameter of S more efficiently?*

- Is there a small *core-set* $Q \subseteq S$ s.t. $\mathrm{diam}(Q(t)) \geq (1 - \varepsilon)\,\mathrm{diam}(S(t))$?

☆ *Kinetic bounding box hierarchies?*

Shape Fitting

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

☆ *Fit a cylinder through S*

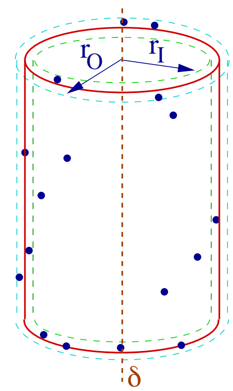
- Find a cylinder C^*

  $C^*(S) = \arg\min_{C} \max_{p \in S} d(p, C)$

☆ For $d = 3$ [A., Aronov, Sharir]

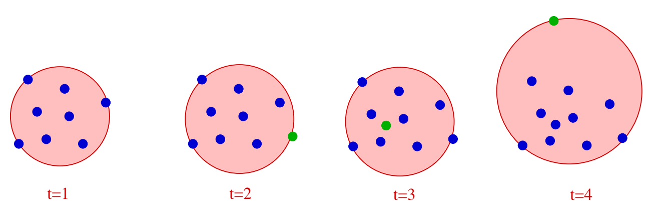
- Optimal solution: n^4
- $O(1)$-approximation: $\approx n^2$

☆ *Can we compute an ε-approximation of $C^*(S)$ in linear time?*



Is there is a small core set *$Q \subseteq S$ so that $C^*(Q)$ approximates $C^*(S)$?*

Geometry in Streaming Model



☆ An incoming stream of points in $\mathbb{R}^d$

☆ Maintain certain statistical measures of the input stream

- Diameter, width, k-clustering

☆ Use $\log^{O(1)} n$ space and processing time

☆ Much work done on maintaining a summary of 1D data

☆ Little known about higher dimensional data
[A., Krishnan, Mustafa, Venkatasubramanian], [Hershberger, Suri], [Bagchi, Chaudhary, Eppstein, Goodrich]

☆ *How much storage and processing time (per point) needed to maintain ε-approximation of* $\mathrm{diam}(S)$*?* Maintain a *core set*!

ε-Approximation and Random Sampling

☆ $X = (S, R)$, $R \subseteq 2^S$: Set system (range space)

- δ: VC-dimension of X

☆ $A \subseteq S$ *ε-approximation* if for all $r \in R$

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

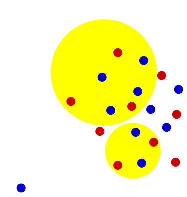
☆ A random subset $A \subset S$ of size $\frac{\delta^2}{\varepsilon^2} \log \frac{\delta}{\varepsilon}$ is an ε-approximation of S with high probability [Vapnik-Chervonenkis]

☆ Efficient deterministic algorithms for computing an ε-approximation [Matoušek, Chazelle]

ε-Approximations

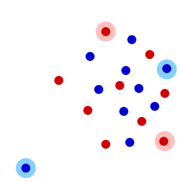
☆ An ε-approximation approximates S in a *combinatorial* sense

- S: Set of points in $\mathbb{R}^2$
- $R = \{r \cap S \mid r \text{ is a disk}\}$
- A: an ε-approximation of (S, R)
- A approximates $|S \cap r|$



☆ A does not approximate 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?

Unified Framework for Core-Sets

☆ Notion of core-set is problem specicific

☆ *Is there a unified framework that constructs core-sets for a wide class of problems?*

- Random subset is an ε-approximation for a large class of range spaces!

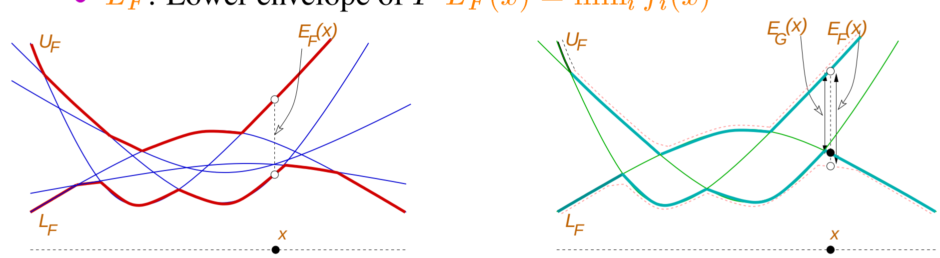
Define the notion of *ε-approximation*

☆ Core set for a wide class of problems

Extents of Functions

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

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



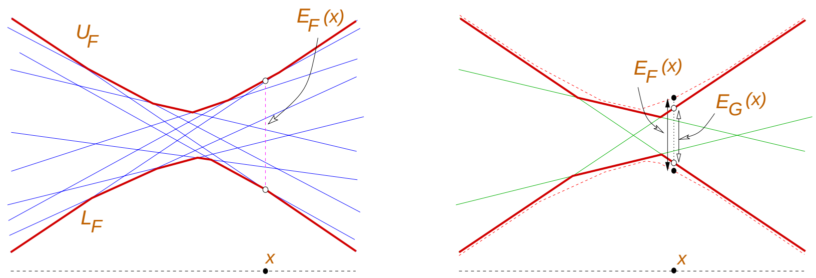
Extent of F:

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

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

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

Linear Functions



- ☆ Many functions can be mapped to linear functions using *linearization*
- ☆ Upper and lower envelopes of linear functions are convex polyhedra
- ☆ Relationship between linear functions and points

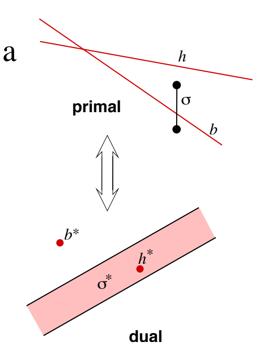
Duality

H: Set of d-variate linear functions

- ☆ Duality: Maps a d-variate linear function to a point in $\mathbb{R}^{d+1}$ and vice-versa

$$h : x_{d+1} = a_1x_1 + \cdots + a_dx_d + a_{d+1}$$
$$\Updownarrow$$
$$h^* = (a_1, \ldots, a_{d+1})$$

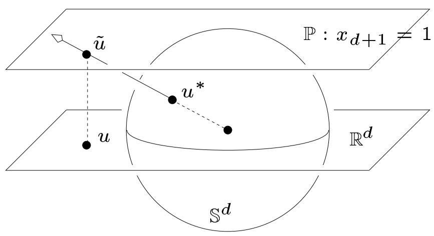
$H^* = \{h^* \mid h \in H\}$

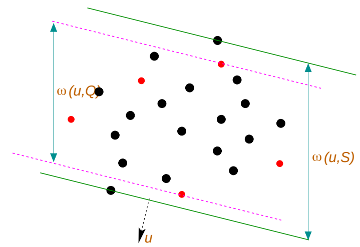


- ☆ *Duality preserves vertical distances.*
- ☆ *Points in $\mathbb{R}^d$ in the primal space map to directions in $\mathbb{S}^d$ in the dual space.*

Duality and ε-Approximations

S: Set of points in $\mathbb{R}^{d+1}$





For $x \in \mathbb{R}^{d+1}$, $\overline{\omega}(x, S) = \max_{p \in S} \langle x, p \rangle - \min_{p \in S} \langle x, p \rangle$

Directional width: For $u \in \mathbb{R}^d$, $\omega(u, S) = \overline{\omega}(\tilde{u}, S)$

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

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

Claim: *$K \subseteq H$ is an ε-approximation of H iff $K^* \subseteq H^*$ is an ε-approximation of H^**

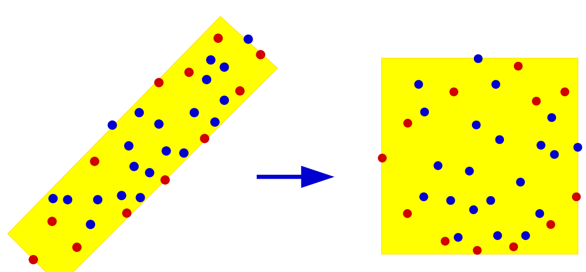
Computing ε-Approximations

Theorem A: *$S \subseteq \mathbb{R}^{d+1}$, $\varepsilon > 0$. We can compute an ε-approximation of S of size*

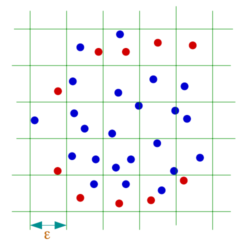
- ☆ $1/\varepsilon^d$ *in time* $n + 1/\varepsilon^d$
- ☆ $1/\varepsilon^{d/2}$ *in time* $n + 1/\varepsilon^{3d/2}$

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

- ☆ $M(S) \in [-1, +1]^{d+1}$, $\mathrm{conv}(M(S))$ is fat
- ☆ Q is an ε-approximation of $S \Leftrightarrow M(Q)$ is an ε-approximation of $M(S)$



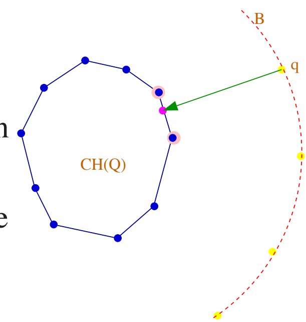
Computing ε-Approximations



Lemma 2: S: Set of n fat points $[-1, +1]^{d+1}$, $\varepsilon > 0$. We can compute an ε-approximation of S of size

- ☆ $1/\varepsilon^d$ in time $n + 1/\varepsilon^d$
- ☆ $1/\varepsilon^{d/2}$ in time $n + 1/\varepsilon^{3d/2}$

Sketch:

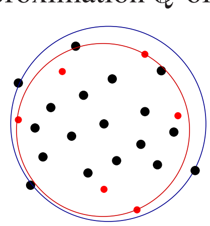


- ☆ Compute $1/\varepsilon^d$-size approximation Q
- ☆ Draw a sphere B of radius 2 centered at origin
- ☆ Draw a grid of size $1/\varepsilon^{d/2}$ on B
- ☆ For each grid point q, select the vertices of the face of $\mathrm{conv}(Q)$ nearest to q

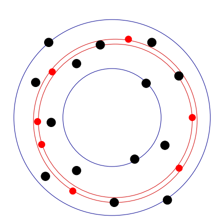
Faithful Extent Measures

$\mu(\cdot)$: Function defined over point sets in $\mathbb{R}^d$ is *faithful* if

- ☆ $\mu(S) \geq 0$ for all $S \subseteq \mathbb{R}^d$
- ☆ $\exists c > 0 \;\; (1 - c\varepsilon)\mu(S) \leq \mu(Q) \leq \mu(S)$ for any ε-approximation Q of S



faithful measure



unfaithful measure

Faithful measures: Diameter, width, radius of smallest enclosing ball, volume of the smallest enclosing box (simplex)

Nonfaithful measures: width of the thinnest spherical shell containing S

Computing Faithful Measures

- ☆ S: Set of points, μ: A faithful measure, $\varepsilon > 0$
- ☆ Compute an (ε/c)-approximation Q of S
- ☆ Compute $\mu(Q)$ using a known algorithm
- ☆ Return $\mu(Q)$
  By definition, $\mu(Q) \geq (1 - \varepsilon)\mu(S)$
- ☆ $S \subseteq \mathbb{R}^d$, $\varepsilon > 0$
  Can compute a pair $p, q \in S$ s.t. $d(p, q) \geq (1 - \varepsilon)\,\mathrm{diam}(S)$
  in time $n + 1/\varepsilon^{3(d-1)/2}$
- ☆ $S \subseteq \mathbb{R}^3$, $\varepsilon > 0$
  Can compute an ε-approximation of the smallest simplex enclosing S
  in time $n + 1/\varepsilon^{9/2}$

ε-Approximations of Linear Functions

Theoream A + Duality:

Theorem B: *H: set of n d-variate linear functions, $\varepsilon > 0$. We can compute an ε-approximation of H of size*

- ☆ $1/\varepsilon^d$ *in time* $n + 1/\varepsilon^d$
- ☆ $1/\varepsilon^{d/2}$ *in time* $n + 1/\varepsilon^{3d/2}$

ε-Approximations of Polynomials

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

Linearization [Yao-Yao, A.-Matoušek]

- ☆ Map $\varphi(x) : \mathbb{R}^d \to \mathbb{R}^k$, $\varphi(x) = (\varphi_1(x), \ldots, \varphi_k(x))$
- ☆ Each f_i maps to a k-variate linear function h_i
- ☆ k: Dimension of linearization

Example: Lifting transform

- ☆ $f(x_1, x_2) = a_3^2 - (x_1 - a_1)^2 - (x_2 - a_2)^2$
- ☆ $\varphi(x_1, x_2) = (x_1, x_2, x_1^2 + x_2^2)$
- ☆ $h(y_1, y_2, y_3) = (a_3^2 - a_1^2 - a_2^2) + 2a_1y_1 + 2a_2y_2 - y_3$

ε-Approximations of Polynomials

Lemma: *$K \subseteq H$ is an ε-approximation of $H \Leftrightarrow$ $G = \{f_i \mid h_i \in K\}$ is an ε-approximation of F.*

Theorem C: *F: a family of n d-variate polynomials, k: dimension of linearization, $\varepsilon > 0$. We can compute an ε-approximation of F of size*

- ☆ $1/\varepsilon^k$ *in time* $n + 1/\varepsilon^k$
- ☆ $1/\varepsilon^{k/2}$ *in time* $n + 1/\varepsilon^{3k/2}$
- ☆ $1/\varepsilon^{\sigma}$ *in time* $n + 1/\varepsilon^{3k/2}$, $\sigma = \min\{d, k/2\}$

Application I: Kinetic Geometry

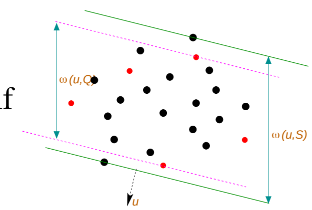
S: Set of n moving points in $\mathbb{R}^d$

- $p_i = a_i + b_i t, \qquad a_i, b_i \in \mathbb{R}^d$
- $S(t) = \{p_i(t) \mid 1 \le i \le n\}$

☆ $Q \subseteq S$ an *ε-approximation* if $\forall u \in \mathbb{R}^{d-1}, t \in \mathbb{R}$

$$(1-\varepsilon)\omega(u, S(t)) \le \omega(u, Q(t))$$

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



Define $f_i(u,t) = \langle p_i(t), \tilde{u} \rangle$; f_i is a $\deg(2)$ polynomial

Claim: $F = \{f_1, \ldots, f_n\}, \qquad \omega(u, S(t)) = E_F(u,t)$

Suffices to compute an ε-approximation of F.

Application I: Kinetic Geometry

Corollary: *S: n moving points in $\mathbb{R}^d$, $\varepsilon > 0$. An ε-approximation of size $1/\varepsilon^{d-1/2}$ can be computed in $n + 1/\varepsilon^{3(d-1/2)}$ time.*

Maintaining the ε-approximate diameter of S:

☆ Compute an ε-approximation Q of S

☆ Use a kinetic data structure to maintain $\mathrm{diam}(Q)$

☆ For $d = 2$

- # events $\approx 1/\varepsilon^3$
- Time spent at each event: $\log(1/\varepsilon)$

☆ Works for maintaining width, smallest enclosing ball/rectangle/simplex, ...

ε-Approximations of Fractional Polynomials

Functions are not polynomials in many applications

- $f_i(x) = d(x, p_i) - r_i$

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

☆ $f_i \equiv (h_i)^{1/r}$, h_i: d-variate polynomial, $r \geq 1 \in \mathbb{N}$

☆ $H = \{h_i \mid 1 \leq i \leq n\}$

Theorem D: *$K \subseteq H$ is an $c\varepsilon^r$-approximation of H, $c > 0$ a constant, then $\{f_i \mid h_i \in K\}$ is an ε-approximation of F.*

Corollary: If H admits a linearization of dimension k, then we can compute an ε-approximation of F of size

☆ $1/\varepsilon^{rk}$ in time $n + 1/\varepsilon^{rk}$

☆ $1/\varepsilon^{r\sigma}$ in time $n + 1/\varepsilon^{3rk/2}$, $\sigma = \min\{d, k/2\}$

Application II: Shape Fitting

S: Set of n points in $\mathbb{R}^2$

☆ *Find the minimum-width annulus containing S.*

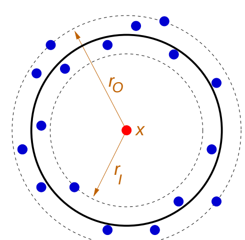
$\mu(x)$: Min width of annulus containing S centered at x

☆ $d(x, p)$: Distance between x and p

$\mu(x) = \max_{p \in S} d(x, p) - \min_{p \in S} d(x, p)$

☆ $f_i(x) = d(x, p_i)$, $F = \{f_1, \ldots f_n\}$

$\mu(x) = E_F(x)$



Compute $w^* = \min_x E_F(x)$

☆ Compute an ε-approximation G of F; $|G| = 1/\varepsilon$

☆ Compute $\boxed{x^* = \arg\min_x E_G(x)}$

☆ Return $E_F(x^*)$; $E_F(x^*) \leq (1 + \varepsilon) w^*$

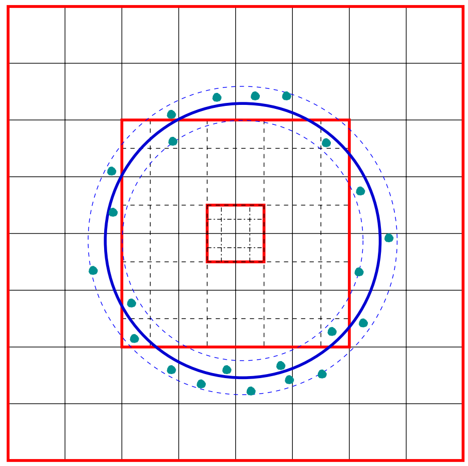
☆ Time: $n + 1/\varepsilon^{O(1)}$

Core-Set for Annulus

☆ Draw an exponential grid on the plane of size $O(1/\varepsilon)$

☆ For each grid cell:

- Choose its center c
- Add the nearest and farthest neighbor of c to the core set



Fitting a Cylinder

S: Set of n points in $\mathbb{R}^3$

☆ *Find the minimum-width cylindrical shell that contains S.*

$\mu(\ell)$: Min width of a shell containing S with axis ℓ

☆ $d(\ell, p)$: Distance between ℓ and p

$\mu(\ell) = \max_{p\in S} d(\ell, p) - \min_{p\in S} d(\ell, p)$

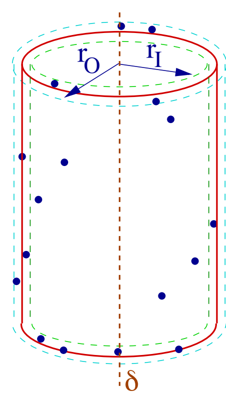
☆ $f_i(\ell) = d(\ell, p_i)$, $F = \{f_1, \ldots f_n\}$

$\mu(\ell) = E_F(\ell)$

☆ Compute $\boxed{w^* = \min_\ell E_F(\ell)}$

- Compute an ε-approximation G of F
- Compute $\ell^* = \arg\min_\ell E_G(\ell)$
- Return $E_F(\ell^*)$; $E_F(\ell^*) \leq (1+\varepsilon)w^*$

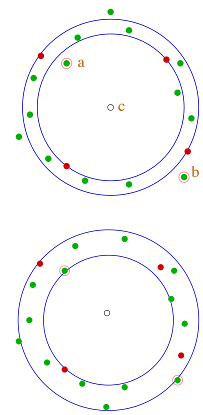
☆ Argue that $(f_i)^2$ is a polynomial



Shape Fitting: Incremental Algorithm

[Varadarajan]

- ☆ S: Set of points in $\mathbb{R}^2$
- ☆ *Find the smallest annulus containing S*
- ☆ A simple iterative algorithm
- ☆ $A \subseteq S$: Initially, $|A| = 4$
- ☆ $W(A)$: Min-width annulus containing A
- ☆ `while` $S \not\subset (1+\varepsilon)W$
  - c: Center of w
  - $a \in S$: Nearest neighbor of c
  - $b \in S$: Farthest neighbor of c
  - $A = A \cup \{a, b\}$



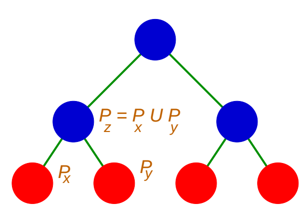
Claim: *The algorithm terminates in $O(1/\varepsilon)$ steps.*

Works for other shape-fitting problems as well.

Dynamization

Maintain an ε-approximation of $S \subseteq \mathbb{R}^{d+1}$ under insertion/deletion

- ☆ Build a balanced-tree T on S
  - h: Height of T
- ☆ Each leaf stores $\approx \left(\frac{h}{\varepsilon}\right)^{d/2}$ points
- ☆ Q_v: $(\varepsilon/2h)$-approximation of $Q_w \cup Q_z$
  - i: height of node v
  - Q_v: $(i\varepsilon/2h)$-approximation of P_v
- ☆ Q_{root} is an $\varepsilon/2$-approximation of S
  - $|Q_{\text{root}}|$: $(h/\varepsilon)^{d/2}$



Maintain an $(\varepsilon/3)$-approximation Q of Q_{root} of size $1/\varepsilon^{d/2}$

Dynamization

Deleting a point p:

- ☆ Find the leaf z that contains p
- ☆ Delete p from z
- ☆ Recompute the ε-approximations at the ancestors of z
- ☆ Update the structure of T if necessary

$$\text{Deletion time: } \left(\frac{\log n}{\varepsilon}\right)^{3d/2} \log n$$

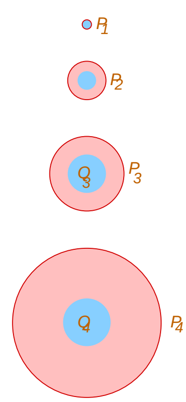
Insertions can be handled similarly

Drawback: Update algorithm is highly *nonrobust*!

Application III: Handling Data Stream

- ☆ S: Stream of points in $\mathbb{R}^2$
- ☆ *Maintain the ε-approximation using $\log^{O(1)} n$ space*
- ☆ Partition P into subsets $P_1, \ldots, P_u$
  - $|P_i| = 2^j$ for some $j \leq \log_2 n$, $j = \text{rank}(P_i)$
  - P_i's are not maintained explicitly
- ☆ Maintain an $(\varepsilon/2)$-approximation Q_i of P_i
  - $|Q_i| = j/\sqrt{\varepsilon}$
  - $\bigcup_i Q_i$ is an $(\varepsilon/2)$-approximation of P.
- ☆ Maintain an $\varepsilon/3$-approximation Q of $\bigcup_i Q_i$

$$|Q| = 1/\sqrt{\varepsilon}$$

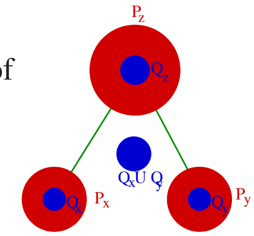


Inserting a Point

☆ Create a new set $P_0 = \{p\}$; $Q_0 = P_0$

☆ If there are two sets P_x, P_y of rank j

- Compute an $\varepsilon/(j+1)^2$-approximation Q_z of $Q_x \cup Q_y$
- Delete Q_x, Q_y and add Q_z;
- $P_z = P_x \cup P_y$; $\text{rank}(P_z) = j + 1$



☆ Q_z is an $(\varepsilon/2)$-approximation of P_z

Space: $\log(n)/\sqrt{\varepsilon}$, Processing time: $\log^3 n/\sqrt{\varepsilon} + 1/\varepsilon^{3/2}$

Corollary: $(1-\varepsilon)$-approximation of $\text{diam}(S)$, $\omega(S)$ can be maintained using $\log(n)/\sqrt{\varepsilon}$ space and $\log^3 n/\sqrt{\varepsilon}$ time.

Also works for

☆ smallest enclosing ball/rectangle/triangle, minimum width annulus, ...

☆ Higher dimensions

Extensions

☆ Computing ε-approximations 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 $1/\varepsilon^{O(1)}$

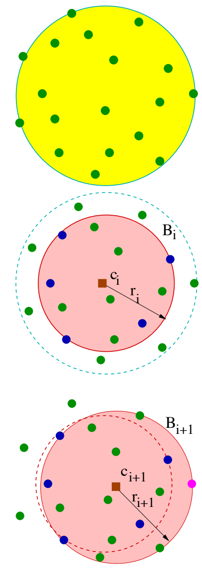
☆ Computing ε-approximations in presence of outliers [Har-Peled, Wang]

☆ Computing ε-approximations for k-clusters
[Har-Peled], [A., Procopiuc, Varadarajan]

- k-centers
- k-line-centers

Minimum Enclosing Balls

[Bădoiu, Clarkson]



☆ S: Set of points in $\mathbb{R}^d$

☆ $C_0 = \{p_i\}$

☆ `repeat` $k = \lceil 2/\varepsilon \rceil$ times

- B_i: Smallest ball enclosing C_i
- c_i, r_i: Center and radius of B_i
- p_{i+1}: Farthest point from c_i
- $\boxed{C_{i+1} = C_i \cup \{p_{i+1}\}}$

☆ Return C_u

☆ R: Radius of the smallest ball enclosing S

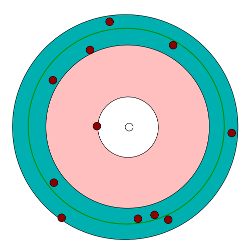
☆ $\lambda_i = r_i/R$

Claim: $\lambda_{i+1} = (1 + \lambda_i^2)/2$

Handling Outliers

☆ P: n points, k: # of outliers, $\varepsilon > 0$

☆ ω_{opt}: width of min-width annulus contains $n - k$ points from P



☆ Find an annulus

- contains $\geq n - k$ points of P
- in time $O(n + (\frac{k}{\varepsilon})^{O(d)})$
- width $\leq (1 + \varepsilon)\omega_{opt}$ — ε-approx.

Key component:

☆ There exists a ε**-coreset** for various fitting problems

- $S \subseteq P$, $|S| = k/\varepsilon^{O(d)}$
- can be computed in linear time
- measure for S ε-approximates that for P

Conclusions

- ☆ ε-approximations in high dimensions
  - Polynomial dependence on $d, 1/\varepsilon$
- ☆ General technique for computing core sets for clustering
- ☆ Core sets for shape fitting if we want to minimize the rms distance
  - Given S, compute a cylinder C so that the rms distance between C and S is minimum
- ☆ Core sets and range spaces with finite VC dimensions

References

- ☆ A., S. Har-Peled, K. Varadarajan, Approximating extent measures of points, *J. ACM*, to appear.
- ☆ M. Bădoiu, S. Har-Peled, P. Indyk, Approximate clustering via core-sets, *34th ACM Sympos. Theory of Computing*, 2002.
- ☆ M. Bădoiu and K. Clarkson, Smaller core-sets for balls, *14th ACM-SIAM Sympos. Discrete Algorithms*, 2003
- ☆ S. Har-Peled and K. Varadarajan, High-dimensional shape fitting in linear time, *19th Annual Sympos. Computational Geometry*, 2003.
- ☆ S. Har-Peled and Y. Wang, Shape fitting with outliers, *19th Annual Sympos. Computational Geometry*, 2003.
- ☆ A., C. M. Procopiuc, and K. Varadarajan, Approximation algorithms for k-line center, *10th Annual European Sympos. Algorithms*, 2002.
- ☆ P. Kumar and E. A. Yildirim, Approximate minimum volume enclosing ellipsoids using core sets, manuscript.