Faster Core-Set Construction and Data-Stream Algorithm in Fixed Dimentions

Timothy M. Chan

Main Results:

1) Diameter: $O(n+1/\varepsilon^{d-3/2})$, improvement from $O(n+1/\varepsilon^{d-1/2})$

2) Width: $O(n+1/\varepsilon^{d-1})$ improvement from $O(n+1/\varepsilon^{3(d-1)/2})$

3) Enclosing Cylinder: $O(n+1/\varepsilon^{d-1})$ improvement from $O(n+1/\varepsilon^{3(d-1)/2})$

4) Dynamic Data structures with smaller polylogarithmic factor for update.

Main Observation:

Let [E] represent integers {1,...,E}

Let p_i denote i-th coordinate of a given point p in d dimentional space.

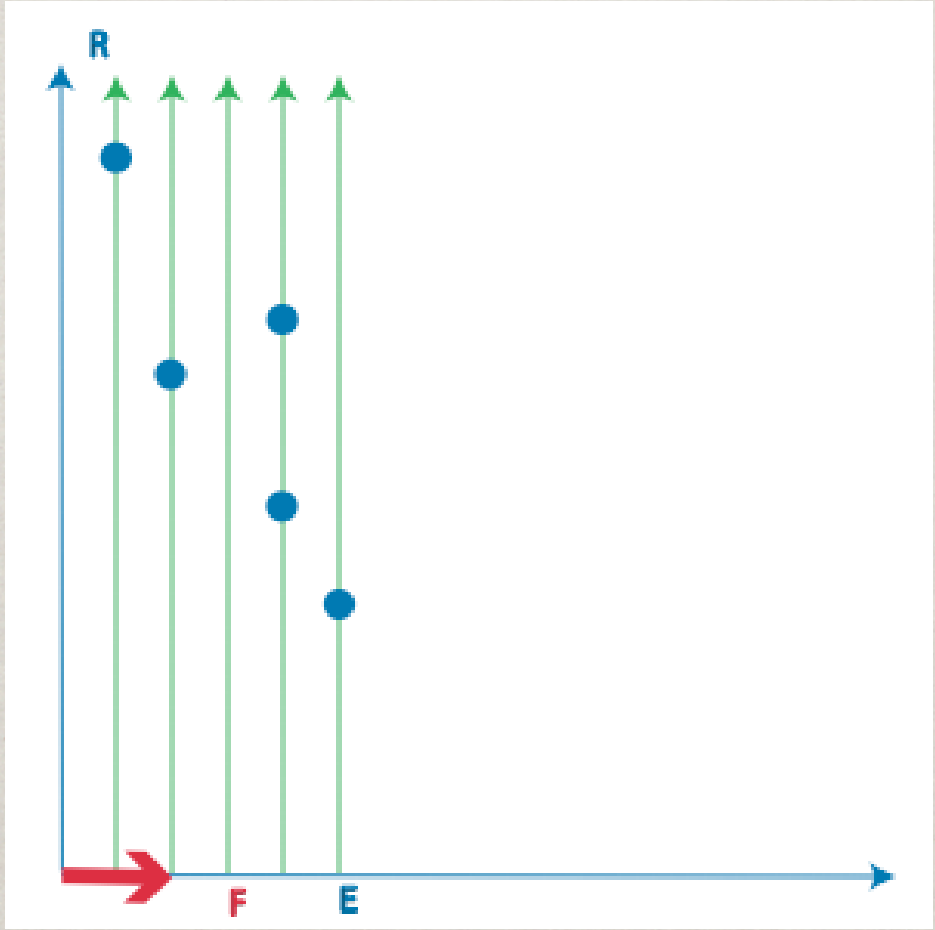
Fix $\delta>0$ and suppose $E^{\delta} \leq F \leq E$, where E and F are integers.

Let $P \subseteq [E]^{d-1} \times \mathbb{R}$ be an n-point set

For every $\xi \in [F]^{d-1}$ compute $q[\xi]=\text{argmax}_{p \in P}(p_1\xi_1 + ... + p_{d-1}\xi_{d-1} + p_d)$

This can be done in $O(n+E^{d-2}F)$ total time.

Geometric Interpretation in 2D

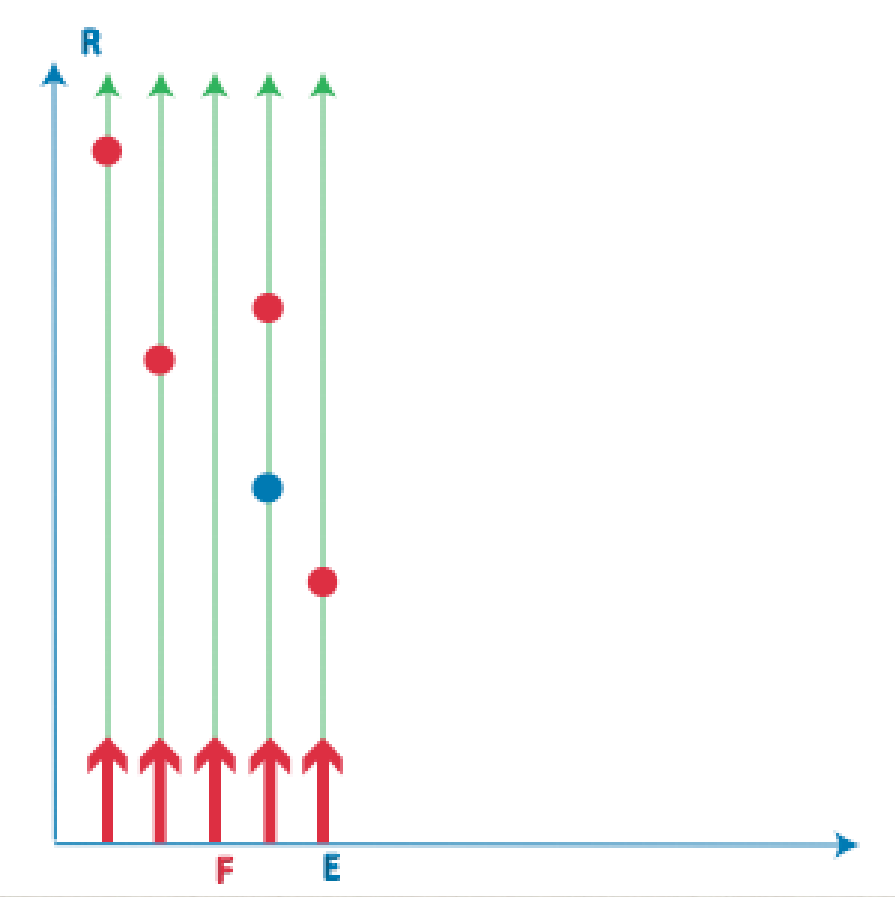


Algorithms

1. for $i \in [E]$ do
2. for $\xi_2, \ldots, \xi_{d-1} \in [F]$ do
3. $r[i, \xi_2, \ldots, \xi_{d-1}] = \text{argmax}_{p_1 = i, p \in P}(p_2\xi_2 + \ldots + p_{d-1}\xi_{d-1} + p_d)$
4. for $\xi_2, \ldots, \xi_{d-1} \in [F]$ do
5. for $\xi_1 \in [F]$ do
6. $q[\xi_1, \ldots, \xi_{d-1}] = \text{argmax}_p(p_1\xi_1 + \ldots + p_{d-1}\xi_{d-1} + p_d)$

   where $p \in \{r[i, \xi_1, \ldots, \xi_{d-1}] \mid i \in [E]\}$

Geometric Interpretation in 2D



Running Time

1. for $i \in [E]$ do
2. for $\xi_2,...,\xi_{d-1} \in [F]$ do
3. $r[i,\xi_2,...,\xi_{d-1}] = \text{argmax}_{p_1=i, p\in P}(p_2\xi_2 + ... + p_{d-1}\xi_{d-1} + p_d)$
4. for $\xi_2,...,\xi_{d-1} \in [F]$ do
5. for $\xi_1 \in [F]$ do
6. $q[\xi_1,...,\xi_{d-1}] = \text{argmax}_p(p_1\xi_1 + ... + p_{d-1}\xi_{d-1} + p_d)$

   where $p \in \{r[i,\xi_1,...,\xi_{d-1}] \mid i \in [E]\}$

Lines 2, 3 is d-1 dimentional subproblem of the same type.

Line 4 is of size F^{d-2}

Line 5, 6 is 2 dimentional subproblem of the same type.

We get running time $T_d(n) = \sum_{i=1}^{E} T_{d-1}(n_i) + F^{d-2}T_2(E) + O(EF^{d-2} + F^{d-1})$

Base case d = 2 can be handeled by explisidly constructing convex hull using radix-sort $O(n + E^{\delta})$ and Grahn's scan. Hence $T_2(n) = O(n+F)$.

By induction and since $F \leq E$ we get $T_d(n) = O(n+E^{d-2}F)$

Corollary

Given $[E]^k \times \mathbb{R}^{d-k}$ we cancompute nearest neighbor to each grid point is $[F]^k \times \{0\}^{d-k}$ it total time $O(n+E^{k-1}F)$

Proof:

Given $\xi \in [F]^k \times \{0\}^{d-k}$, $\operatorname{argmin}_{p \in P} \|p-\xi\|$ is also minimizing $\|\xi\|^2 - 2p_1\xi_1 + \ldots + 2p_k\xi_k + \|p\|^2$. Which is the same as to find $\operatorname{argmax}_{p \in P}(2p_1\xi_1 + \ldots + 2p_k\xi_k - \|p\|^2)$. $p_{k+1} + \ldots + p_d = p' \in \mathbb{R}$,

Diameter

Theorem 1:

Suppose origin $o \in$ box B, where the boundary ∂B is of distance at least 1 from o. Given an ε-grid over ∂B, for any vector x $\exists$ a grid point ξ such that the angle $\angle(\xi,x)$ between $o\xi$ and x is at most $\arccos(1-\varepsilon^2/8)$

Proof:

By scaling assume that x $\in \partial$B. Clearly $\exists\ \xi : \|\xi - x\| \le \varepsilon/2$.

Hence $2\xi \bullet x \ge \|\xi\|^2 + \|x\|^2 - \varepsilon^2/4 \ge 2\|\xi\|\|x\| - \varepsilon^2/4 \ge 2\|\xi\|\|x\|(1-\varepsilon^2/8)$.

Now since $\cos\angle(\xi,x) = \xi \bullet x / \|\xi\|\|x\|$ theorem follows.

Diameter Cont'd

Theorem:

The diameter of n points set in $\mathbb{R}^d$ can be approximated to within a factor of $1+\varepsilon$ in $O(n+1/\varepsilon^{d-3/2})$

Proof:

1) Compute 2 approximation by takin a random point which will be the origin o and finding the point furtherst from it - v. The diameter $\Delta \subseteq (\|v\|, 2\|v\|)$

2) Round each point $p \in P$ to point p' on an $(\varepsilon\Delta)-$grid.

3) Let Ξ be the points of a $\sqrt{\varepsilon}-$grid over $\partial[-1,1]^d$

4) Return the fartherst pair among $\{(p_\xi, q_\xi)\}_{\xi\in\Xi}$ where $p_\xi, q_\xi \in P$ maximize $(p'_\xi - q'_\xi)\bullet\xi$

Diameter Cont'd

Clearly this is the direct reduction to the main observation with $E = O(1/\varepsilon)$ and $F = O(1/\sqrt{\varepsilon})$; since we are maximizing $p' \bullet \xi$ and minimizing $q' \bullet \xi$ for each point on 2d grid of dimention d-1. We get the running time quoted.

Remember that $\angle(\xi, p'-q') \leq \arccos(1-\varepsilon/8)$. Then:

$(p'_\xi - q'_\xi) \bullet \xi \geq (p' - q') \bullet \xi \Rightarrow$ by Cauch-Schwartz

$\left\|p'_\xi - q'_\xi\right\|\|\xi\| \geq \|p' - q'\|\|\xi\|(1 - \varepsilon/8) \Rightarrow$ by definition of rounding

$\left\|p_\xi - q_\xi\right\| + \varepsilon\Delta \geq (\|p - q\| + \varepsilon\Delta)(1 - \varepsilon/8)$

since $\max_{\xi \in \Xi} \left\|p_\xi - q_\xi\right\|$ is $1 + O(\varepsilon)$ approximation of diameter we can readjust ε to get approximation factor of $1+\varepsilon$

Dynamic Cylinder Approx.

Let Rad(P) denote minimum radius of all cylinders enclosing point set P

Let d(p, l) denote the distance between point p and line l.

WLOG let o be the origin.

Let v be the point furtherst from o.

Theorem1:

$$d(p, ov) \leq 2\left(\frac{\|p\|}{\|v\|}+1\right)\mathrm{Rad}(\{o,v,p\})$$

Proof:

In the triangle $\triangle ovp$ let h_1, h_2, h_3 be the respective altitudes for ov, op and pv.

$$2\mathrm{Rad}(\{o,v,p\}) = \min\{h_1,h_2,h_3\} = h_1 \min\left\{1, \frac{\|v\|}{\|p\|}, \frac{\|v\|}{\|p-v\|}\right\}, \text{ since } h_1 = d(p,ov)$$

$$h_1 \min\left\{1, \frac{\|v\|}{\|p\|}, \frac{\|v\|}{\|p-v\|}\right\} \geq d(p,ov)\left(\frac{\|v\|}{\|p\|+\|v\|}\right).$$

Dynamic Cylinder Approx. Cont'd

Theorem2:

Given a stream of points in $\mathbb{R}^d$ we can maintain factor-18 approximation of minimum radius over all enclosing cylinders in a single pass and O(d) space and update time.

Proof:

1) start with two points o,v, set value of w = 0.

insert(p):

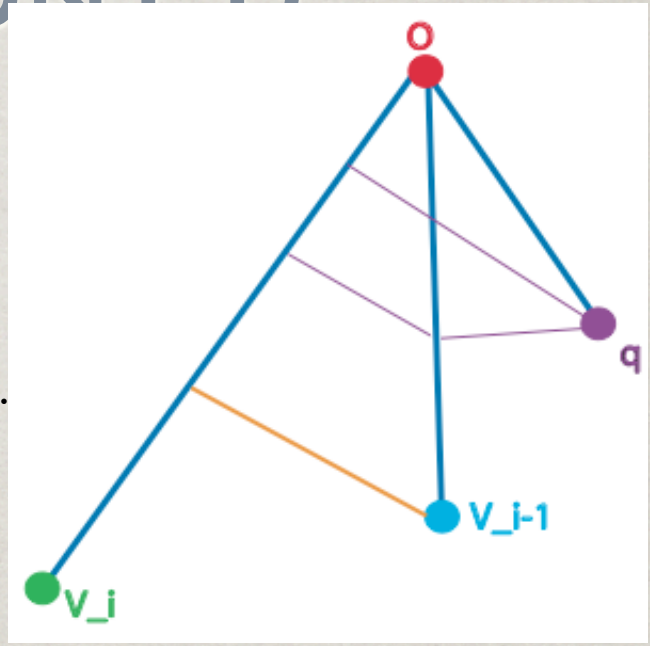
2) w = max{w, Rad({o,v,p})}

3) if $\|p\| > 2\|v\|$ then v = p

Let w_f and v_f refer to the final values of w and v, and let v_i refer to the value after insertion i.

Remember that $\|v_i\| > 2\|v_{i-1}\|$ for all i.

Dynamic Cylinder Approx. Cont'd



Assume that we just added point q.

By theorem 1: $d(q, ov_j) \leq 2(2+1)\text{Rad}(\{o,v_j,q\}) \leq 6w_f$.

For every $i > j$: $d(q, ov_i) \leq d(q, ov_{i-1}) + d(q', ov_i)$
where q' is the orthoganal projection of q on ov_{i-1}.

By similarity of triangles: $d(q', ov_i) = (\|q'\| / \|v_{i-1}\|) d(v_{i-1}, ov_i)$
and $d(q', ov_i) \leq (\|q\| / \|v_{i-1}\|) 3w_f$

Therefore:
$d(q, ov_i) \leq d(q, ov_{i-1}) + 3w_f \|q\| / \|v_{i-1}\|$

Dynamic Cylinder Approx. Cont'd

We had $d(q, ov_i) \leq d(q, ov_{i-1}) + 3w_f \|q\| / \|v_{i-1}\|$

Now due to the doubling property we get:

$$d(q, ov_f) \leq d(q, ov_j) + 3w_f(1 + 1/2 + 1/4 + \ldots)\|q\| / \|v_j\| \leq 6w_f + 2(2)3w_f = 18w_f$$