Data Stream Algorithms in Computational Geometry

Timothy M. Chan
School of CS
Univ of Waterloo

INTRODUCTION

Data Stream: The Basic Model

makes one pass over input
uses limited (sublinear) space

Motivation

MASSIVE data sets

Brief History [Indyk]

Ancient times

(finite automata, sorting w. few passes,
median [Munro, Paterson '80])

Middle ages

Renaissance

([Alon, Matias, Szegedy '96],
[Henzinger, Raghavan, Rajagoplan '98],
& TONS of papers ...
see Muthukrishan's survey)

Data Stream Meets CG

diameter [Feigenbaum, Kannan, Zhang '02, Hershberger, Suri '03]

other measure problems like width [Agarwal, Har-Peled, Varadarajan '04, Chan (SoCG'04)]

statistical problems like range counting [Bagchi, Chaudhary, Eppstein, Goodrich (SoCG'04) Suri, Toth, Zhou (SoCG'04)]

clustering problems like k-median/k-means [Har-Peled, Mazumdar (STOC'04)]

Euclidean MST/matching [Indyk (STOC'04)]

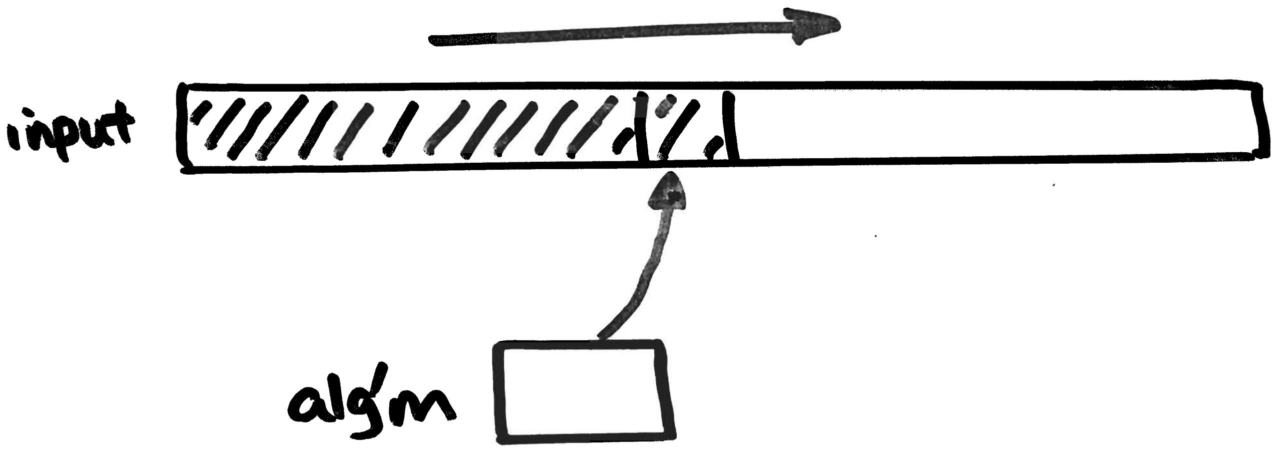
This Talk

looks at some specific examples
& some recent results

[Chan (SoCG'04),
Chan, Sadjad (ISAAC'04),
Chan, Chen (SoCG'05)]

illustrates a few techniques
& different types of
geometric streaming alg/ms

ONE-PASS ALG/MS



(data structures w. sublinear space supporting inserts)

Example 0: Approx Median in 1D

Given n pts in $\mathbb{R}^1$,

find a pt of rank $\sim \frac{n}{2} \pm \varepsilon n$

Munro, Paterson's Alg'm ['80]

Standard idea: the "logarithmic" method
[Bentley, Saxe '80]

Divide set into $O(\log n)$ groups
whose sizes are distinct powers of 2

To insert:
- create new group of size 1
- whenever 2 groups have same size, merge

Modified idea: "merge-and-reduce"

Replace each group with sketch

Def: $R \subset S$ is a δ-sketch of S iff i^{th} pt of R has rank $\sim \frac{in}{r} \pm \delta n$

$(n = |S|,\ r = |R|)$

Fact: $\exists$ δ-sketch of size $1/\delta$

Fact: If R_j is α-sketch of S_j

$(|R_1| = |R_2|,\ |S_1| = |S_2|)$

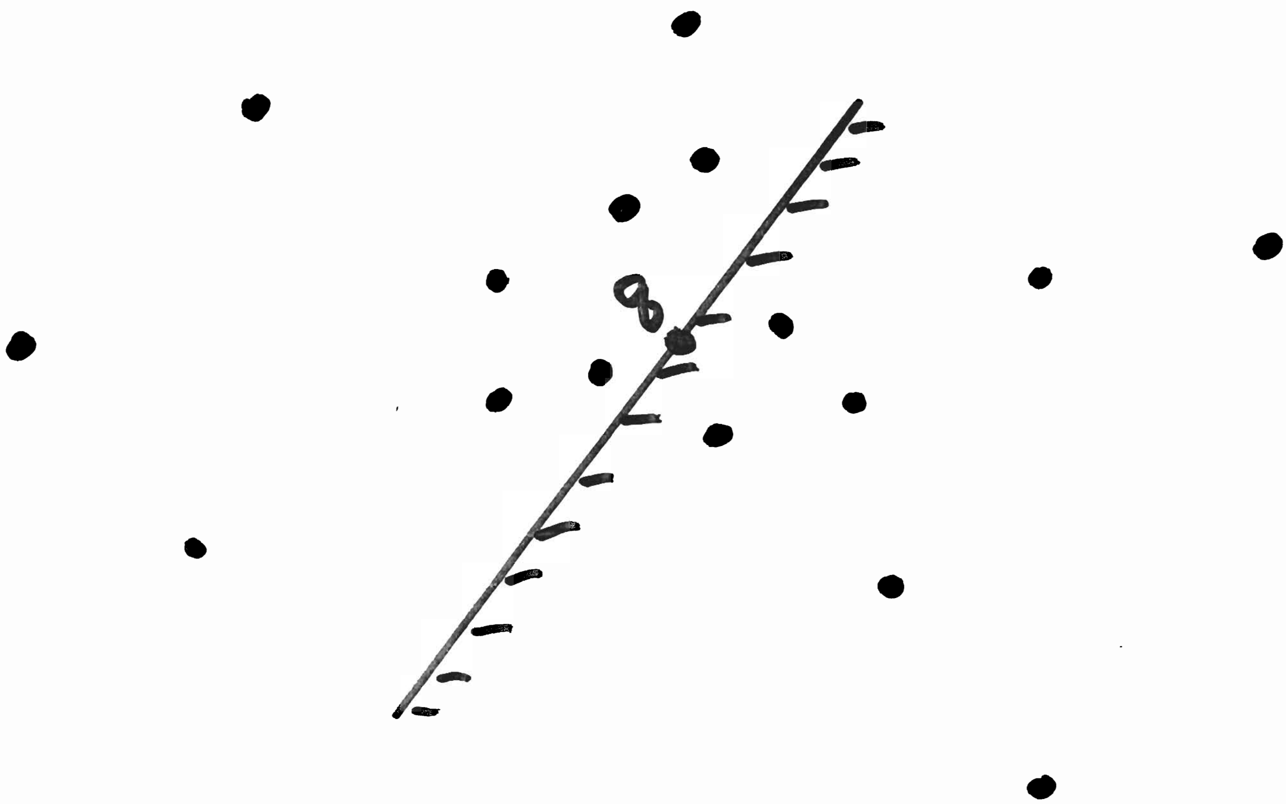
then δ-sketch of $R_1 \cup R_2$ is $(\alpha + \delta)$-sketch of $S_1 \cup S_2$

Set $\delta = \varepsilon / \log n$

$\Rightarrow$ space $O(\frac{1}{\delta} \log n) = \underline{\underline{O(\frac{1}{\varepsilon} \log^2 n)}}$

Example 1 : Approx Centerpoint

Given n pts in $\mathbb{R}^d$,
find a pt $q \in \mathbb{R}^d$ s.t.
any halfspace containing q
contains $\geq \frac{n}{d+1} - \varepsilon n$ pts



Bagchi, Chaudhary, Eppstein, Goodrich's Alg'm ['04]

Idea: same method!

Def: [Vapnik, Chervonenkis '71 ... Matoušek '95]

RCS is a δ-approximation of S iff any halfspace containing i pts of R contains $\sim \frac{in}{r} \pm \delta n$ pts of S

Fact: $\exists$ δ-approximation of size $O(\frac{1}{\delta^2} \log \frac{1}{\delta})$

$\Rightarrow$ space $\underline{\underline{O(\frac{1}{\varepsilon^{O(1)}} \text{polylog } n)}}$

Other Applications [BCEG'04]

more statistics problem

(simplicial depth, LMS, ...)

range counting

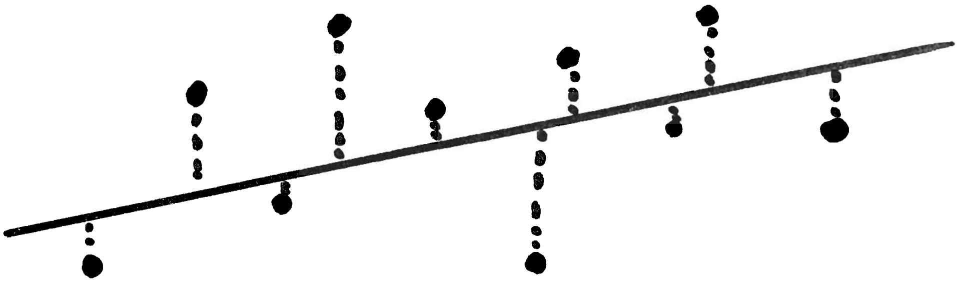
Example 2: Approx Hyperplane Fitting

Given n pts in $\mathbb{R}^d$,

find hyperplane s.t.

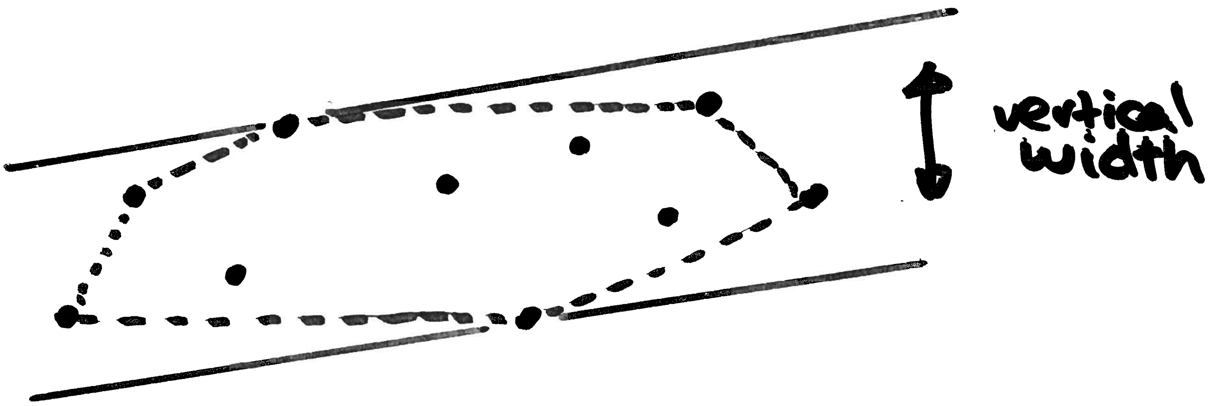
max vertical distance to the pts

is minimized, to within $1+\varepsilon$ factor



i.e. minimize vertical-width of

convex hull (CH) over all directions



Agarwal, Har-Peled, Varadarajan's Alg'm ['04]

Idea: logarithmic method again!

Def: $R \subset S$ is a δ-core-set of S iff
$\forall$ direction x
width of $CH(S)$ along x
$\leq (1+\delta) \cdot$ width of $CH(R)$ along x

Fact: $\exists$ δ-core-set of size $O\left(\frac{1}{\delta^{(d-1)/2}}\right)$

$\Rightarrow$ space $\underline{\underline{O\left(\frac{1}{\varepsilon^{O(1)}} \text{poly} \log n\right)}}$

A New Alg'm [C'04]

Idea: "geometric-series" method

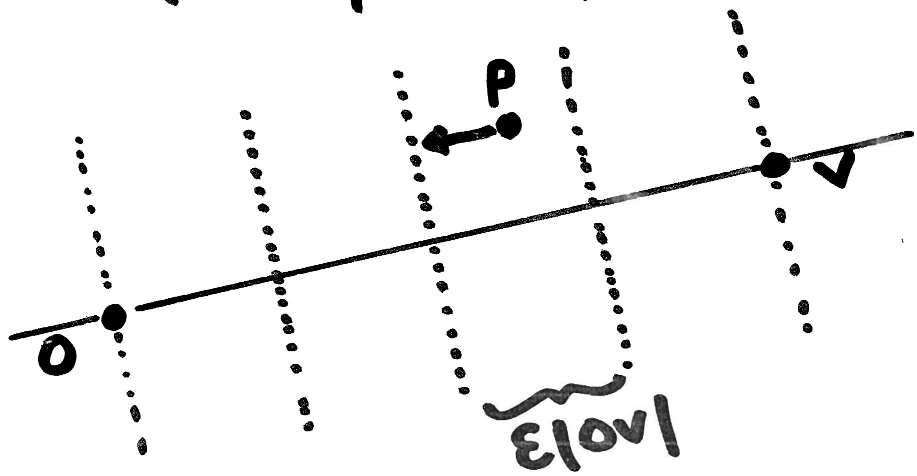
Static version: // core-sets in 2D

o = first pt

v = (approx) farthest pt from o

round pts to $\Theta(1/\varepsilon)$ grid lines orthogonal to $\overleftrightarrow{ov}$

keep min/max pts on each line



Analysis: rounding error

$\leq \varepsilon \cdot$ width of $\overline{ov}$ along x

$\leq \varepsilon \cdot$ width of $CH(S)$ along x

Modified streaming version:

To insert p:

whenever $|op| > 2|ov|$,

$v = p$

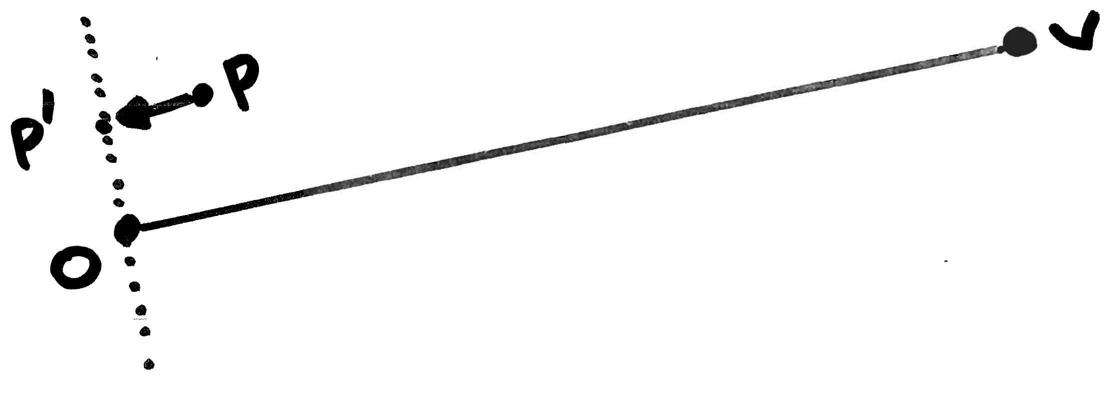
start a new core-set & clean up

To clean up:

keep only the b newest core-sets

round old pts to grid line thru o

keep min/max pt on line



Analysis:

$$\text{rounding error} \leq \frac{|pp'|}{|ov|} \cdot \text{width of } \overline{ov} \text{ along } x$$

$$\leq \frac{|op|}{|ov|} \cdot \quad \text{" " "}$$

when p was inserted: $|op| \leq 2|ov|$

at every change: $|ov|$ doubles

when p gets old: $|op| \leq \frac{2|ov|}{2^b}$

$\Rightarrow$ total error $\leq 2\left(\frac{1}{2^b} + \frac{1}{2^{b+1}} + \dots\right) \cdot$ width of CH(S) along x

Set $b = \log \frac{1}{\varepsilon}$

$\Rightarrow$ space $O\left(b \cdot \frac{1}{\varepsilon}\right) = \underline{\underline{O\left(\frac{1}{\varepsilon} \log \frac{1}{\varepsilon}\right)}}$

const (indep of n)!

Other Results & Applications [C'04]

refinement in 2D:

space $O(\frac{1}{\sqrt{\varepsilon}} \log^2 \frac{1}{\varepsilon})$

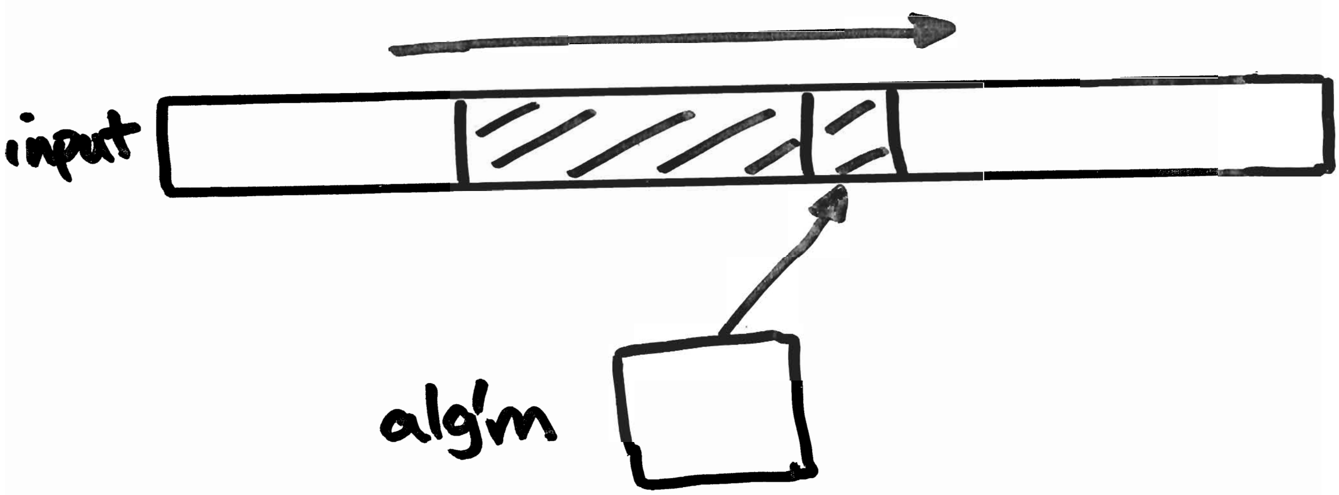
core-sets in dD:

space $O(\frac{1}{\varepsilon^{d-0.5}} \log^d \frac{1}{\varepsilon})$

bounding box

sphere/cylinder fitting

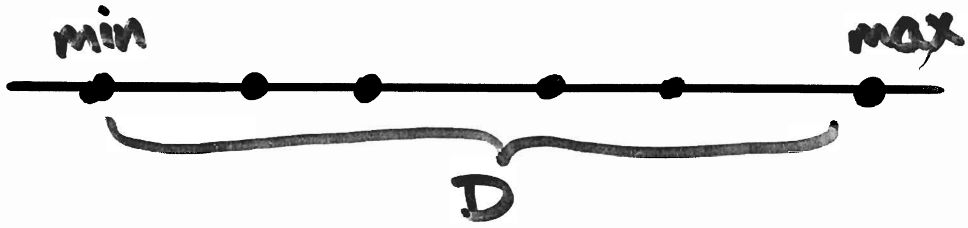
SLIDING-WINDOW ALG/MS



(data structures supporting
"insert" & "delete oldest pt")

Example : Approx Diameter in 1D

Given n pts in $\mathbb{R}^1$,
compute D = farthest distance
to within $1+\varepsilon$ factor

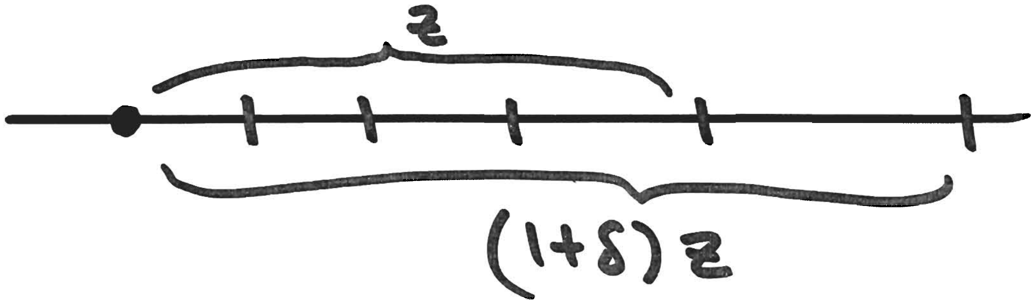


Feigenbaum, Kannan, Zhang's Algm ['02]

Idea: logarithmic method again!

How to sketch:

forms exponentially-spaced grid
keeps newest pt (the representative) of each cell



$\Rightarrow$ space $\underline{\underline{O\left(\frac{1}{\varepsilon} \log \Delta \log^2 n\right)}}$

where $\Delta = \dfrac{\text{farthest dist}}{\text{closest dist}}$

A New (Simple!) Algm [C, Sadjad '04]

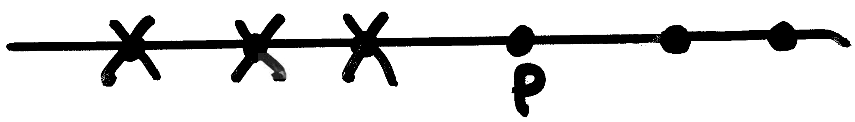
Idea: "skip-middle" method

To insert p: // approx max

insert p to R

remove all pts $< p$ from R

after $\frac{1}{\varepsilon} \log \Delta$ inserts, clean up



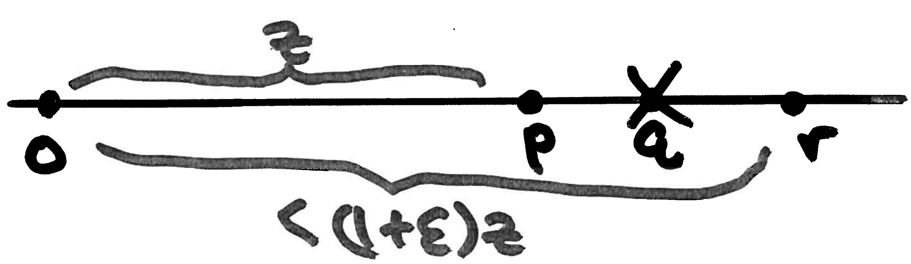
To clean up:

o = smallest pt of R

whenever $\exists$ 3 consecutive $p, q, r \in R$

s.t. $|or| < (1+\varepsilon)|op|$,

delete q from R

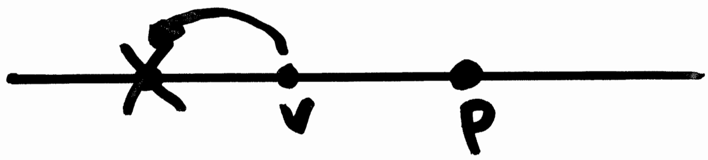


$\Rightarrow$ space $|R| = \underline{\underline{O\left(\frac{1}{\varepsilon} \log \Delta\right)}}$ optimal!

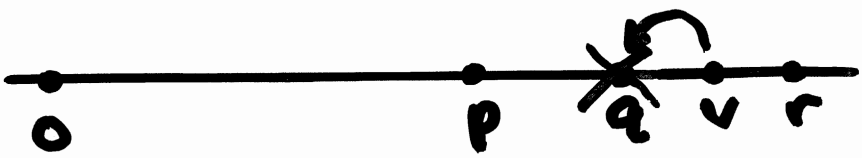
Correctness / Invariants:

1. smaller pts are newer in R
2. each pt v has a representative $v' \in R$ that is newer than v, where either
   - 2.1. v' is successor of v in R, or
   - 2.2. v' is predecessor " " " " with $|vv'| \leq \varepsilon \cdot$ diameter of all pts newer than v

(insert)



(clean-up)



$$\left(|or| \leq (1+\varepsilon)|op| \Rightarrow |pv| \leq \varepsilon |op| \right)$$

Other Applications [CS'04]

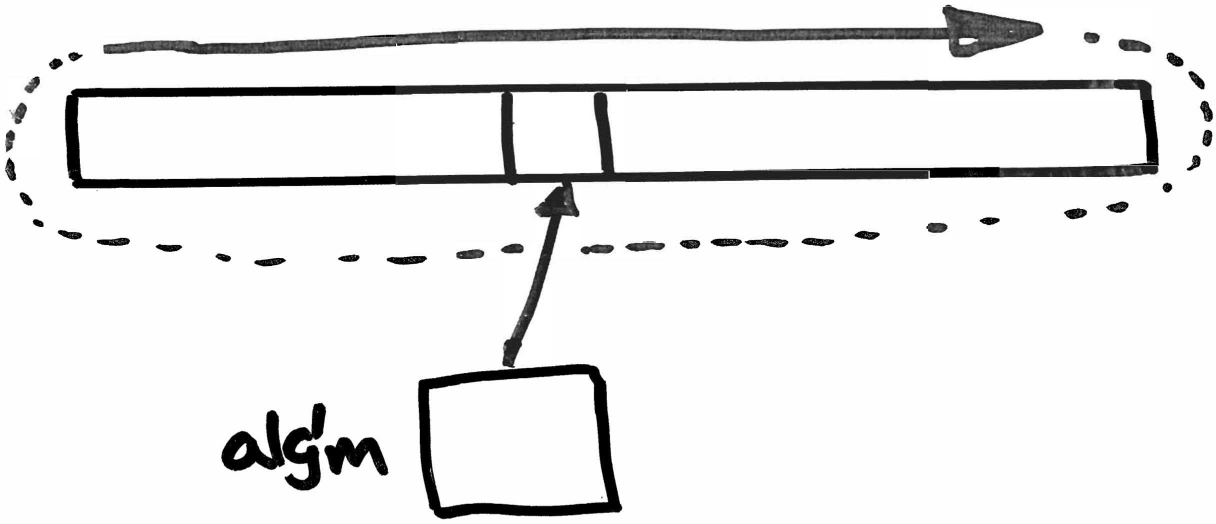
diameter in dD :

space $O\left(\frac{1}{\varepsilon^{(d-1)/2}} \log \frac{\Delta}{\varepsilon}\right)$

[by projection]

core-sets in 2D

MULTI-PASS ALG'MS



1 Pass vs. 2 Passes, according to Knuth [AOCP, vol 1, '73]

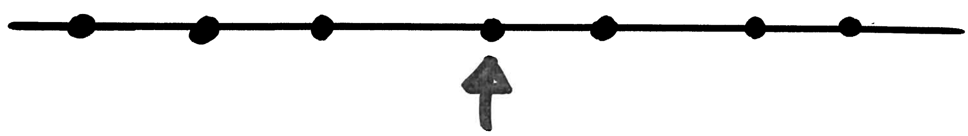
Old lady, on a bus:

"Can you tell me how I can get off Pasadena St.?"

Boy:

"Just watch me, and get off two stops before I do."

Example 0 : Exact (!) Median



Munro, Paterson's Alg/m ['80]

Idea: filtering ("prune-and-search")

$I = (-\infty, \infty)$

repeat

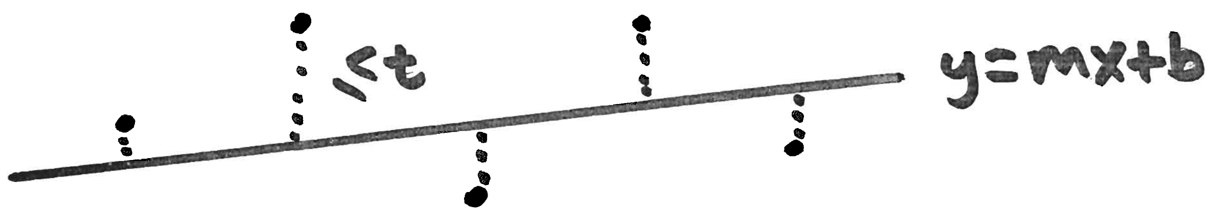
1. among pts inside I,
   compute $(1/r)$-sketch R of size $O(r)$
2. I = sub-interval containing answer

Set $r = n^\delta$

$\Rightarrow$ passes $\log_r n = \underline{\underline{\frac{1}{\delta}}}$ const!

space $O(r \log^2 n) = \tilde{O}(n^\delta)$ small!

Example 2 : Exact Hyperplane Fitting



$$\min t$$

$$\text{s.t.} \quad y_i \le m x_i + b + t$$

$$y_i \ge m x_i + b - t$$

$$(i = 1, \dots, n)$$

reduces to linear programming (LP) in $\mathbb{R}^{d+1}$

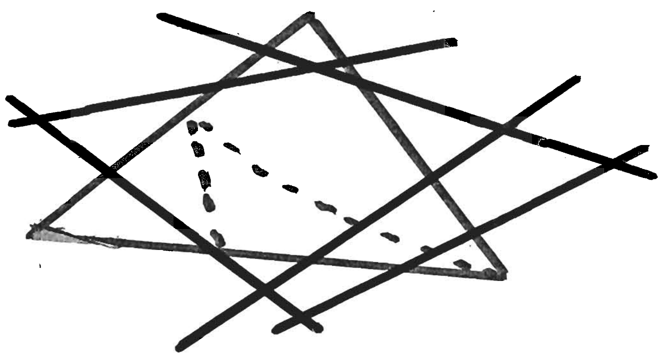
"New" Alg'm [C, Chen '05]

Idea: prune-&-search again

$\Delta = \mathbb{R}^d$ // LP in dD

repeat:

1. among halfspaces crossing Δ, compute $(1/r)$-cutting of size $r^{O(1)}$
2. Δ = sub-simplex containing answer



Def: Given n halfspaces, δ-cutting is a partition of space into simplices each crossed by $\leq \delta n$ halfspaces

Analysis:

1. take $(1/r)$-approximation
2. solve $r^{O(1)}$ LPs in $(d-1)$D
   "in parallel"

passes $P_d(n) = O(P_{d-1}(n) \log_r n)$

space $S_d(n) = O(S_{d-1}(n)\, r^{O(1)} \log_r n)$

Set $r = n^{c\delta}$

$\Rightarrow$ passes $O(1/\delta^{d-1})$ const!

space $\tilde{O}(n^{\delta})$ small!

Other Results [CC'05]

refinement: $O(n)$ rand time

$O(n \log\log\cdots n)$ det time in 2D

lower bd for LP in 2D:

$\frac{1}{\delta}$ passes $\Rightarrow$ $\Omega(n^{\delta})$ space

(as in Munro, Paterson

CH of sorted pts in 2D:

$O(1)$ passes, $O(n^{1/2+\delta})$ space,

$O(n)$ time

CH for small output size h in 2D:

$O(1)$ passes, $O(hn^{\delta})$ space,

$O(n\log n)$ time

CONCLUSION

Summary of New Results

one-pass alg'ms for core-sets
(approx CH) [C'04]

sliding-window alg'ms for
diameter [CS'04]

multi-pass alg'ms for LP [CC'05]

Some Open Problems

one-pass for high dim

e.g. diameter: $\sim\sqrt{2}$ factor [Indyk'03]

min enclos cylinder:

~ 5 factor [C'04]

smallest # passes

e.g. LP : $d+1$ passes, $\tilde{O}(\sqrt{n})$ space

[Clarkson's alg'm]

lower bd ?