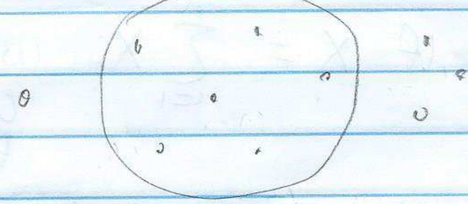
Approx Counting

<u>Problem1</u> Given set P of n pts in $\mathbb{R}^2$, and fixed radius r, find disk of radius r minimizing # pts outside (or maximizing # pts inside)



[known exact alg'ms: $O((n+k^2)\log^{O(1)} n)$]

<u>Problem2</u> build data structure s.t. given any query disk / halfspace / rectangles / ... q, can quickly count # of pts inside / outside q.

("range counting")

idea - ε-coreset for counting?

<u>Def</u> (Vapnik - Chervonenkis '71)

An <u>ε-approximation</u> is a subset $R \subseteq P$ s.t. (of size r)

$\forall$ disk q,

$$\left| \frac{|P \cap q|}{n} - \frac{|R \cap q|}{r} \right| \leq \varepsilon$$

(appl: learning theory ...)

i.e. $|P \cap q| \approx |R \cap q| \cdot \frac{n}{r} \underbrace{\pm\, \varepsilon n}_{\text{additive error}}$

<u>Main Thm</u> $\exists$ ε-approximation of size $O\left(\frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon}\right)$

Pf: just take random sample R !!

("probabilistic method")

(e.g. $\Pr[|X-\mu| \geq c\sigma] \leq \frac{1}{c^2}$) **Chebyshev ineq** If also $Var[X] = \sigma^2$,

$$\Pr[|X-\mu| \geq D] \leq \frac{\sigma^2}{D^2}$$

Pf: $\Pr[(X-\mu)^2 \geq D^2]$

(e.g. $\Pr[X \geq c\mu] \leq \frac{1}{c}$) **Markov's ineq** If $X \geq 0$, with $E[X] = \mu$,

$$\Pr[X \geq t] \leq \frac{\mu}{t}$$

(Pf: $E[X] \geq \sum_{a \geq t} t \Pr(X=a)$)

Chernoff's ineq If $X = \sum_{i=1}^{n} X_i$ where X_i's are indep 0–1 vars with $\Pr[X_i = 1] = p$. ($\mu = np$)

$$\Pr[|X-\mu| \geq D] \leq \frac{1}{e^{\Theta(D^2/\mu)}}$$

(e.g. $\Pr[|X-\mu| \geq c\sqrt{\mu}] \leq \frac{1}{e^{\Theta(c^2)}}$) $(0 \leq D \leq \mu)$

Pf (of one dir):

$$\Pr[X \geq \mu + D]$$
$$= \Pr[b^X \geq b^{\mu+D}]$$
$$\leq \frac{E[b^X]}{b^{\mu+D}} \quad \text{by Markov's ineq}$$
$$= \frac{E[\prod_{i=1}^{n} b^{X_i}]}{b^{\mu+D}} = \frac{\prod_{i=1}^{n} E(b^{X_i})}{b^{\mu+D}} \quad \text{by indep}$$
$$= \frac{(p b^1 + (1-p) b^0)^n}{b^{\mu+D}}$$
$$= \frac{(1+(b-1)p)^n}{b^{\mu+D}} \leq \frac{e^{(b-1)pn}}{b^{\mu+D}} = \left(\frac{e^{b-1}}{b^{1+\delta}}\right)^\mu$$

Let $\delta = D/\mu$ $(0 \leq \delta < 1)$

choose $b = 1+\delta$

$$= \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu$$

$$\simeq \left(\frac{1}{e^{(1+\delta)\underbrace{\ln(1+\delta)}_{\delta-\Theta(\delta^2)}-\delta}}\right)^\mu = \frac{1}{e^{\Theta(\delta^2\mu)}}$$

$$= \frac{1}{e^{\Theta(D^2/\mu)}} \quad \square$$

Back to Pf of Main Thm:

Fix disk q.

Say $R = \{p_1 \dots, p_r\}$ where each p_i is randomly chosen (with replacement)

Let $X_i = \begin{cases} 1 & \text{if } p_i \in q \\ 0 & \text{else} \end{cases}$ $\qquad X = |R \cap q|$

$\Pr[X_i = 1] = \frac{|P \cap q|}{n}$ $\qquad \mu = \frac{|P \cap q|}{n} r$

$\Rightarrow \Pr\left[\left| \frac{|R \cap q|}{r} - \frac{|P \cap q|}{n} \right| > \varepsilon \right]$

$= \Pr[|X - \mu| > \varepsilon r]$

$\leq \frac{1}{e^{\Theta(\varepsilon^2 r^2/\mu)}}$

minor case: $\varepsilon r > \mu$

$\leq \frac{1}{e^{\Theta(\varepsilon^2 r)}}$ $\qquad (\mu \leq r)$

$= \frac{1}{e^{\Theta(b)}}$ $\qquad$ Set $r = \frac{b}{\varepsilon^2}$

There are $O(n^3)$ "different" disks q

$\Rightarrow \Pr[R \text{ fails for some } q]$

$\leq \frac{n^3}{e^{\Theta(b)}} < 1$ for $b =$ large const $* \ln n$

$\Rightarrow \exists$ ε-approximation of size $O(\frac{1}{\varepsilon^2} \log n)$

How to reduce $\log n$ to $\log \frac{1}{\varepsilon}$?

idea - decrease size gradually ...

)

Fact If R is an ε-approximation of P
& R' is an ε'-approximation of R,
then R' is an $(\varepsilon+\varepsilon')$-approximation of P.

Set $\varepsilon = c\sqrt{\frac{\log n}{n}} \Rightarrow \exists\ O\left(\sqrt{\frac{\log n}{n}}\right)$-approximation of size $\frac{n}{2}$

$\Rightarrow \exists\ O\left(\sqrt{\frac{\log n}{n}} + \sqrt{\frac{\log(n/2)}{n/2}} + \dots + \sqrt{\frac{\log(2t)}{2t}}\right)$ - approximation of size t

$\Rightarrow \exists\ O\left(\sqrt{\frac{\log t}{t}}\right)$-approx. of size t

Set $t = \frac{c'}{\varepsilon^2}\log\frac{1}{\varepsilon}$

$\Rightarrow \exists\ \varepsilon$-approx. R of size $O\left(\frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon}\right)$. $\square$

Rmk 1: at the end of above construction, R is still a random sample! (sample of a sample of ...)

Rmk 2: very general ("range space"/"set system")
a collection $\mathcal{Q}$ ("ranges") of subsets of U has shattering dim d (related to "VC dim")
if $\forall P \subseteq U$, $|\{P \cap q : q \in \mathcal{Q}\}| \le O(|P|^d)$

Rmk 3: can be derandomized in polytime (by method of conditional probability)

Rmk 4: related to discrepancy theory

Def P has discrepancy $\le \Delta$ if there is a 2-coloring of P
s.t. $\forall$ disk q,
$$|\#\text{ red pts in } P\cap q - \#\text{ blue pts in } P\cap q| \le \Delta$$

i.e. $|\#\text{ red pts in } P\cap q| \approx \frac{1}{2}|P\cap q| \pm \frac{\Delta}{2}$

discrepancy $\le \Delta \Rightarrow O(\frac{\Delta}{n})$ - approx. of size $\sim \frac{n}{2}$

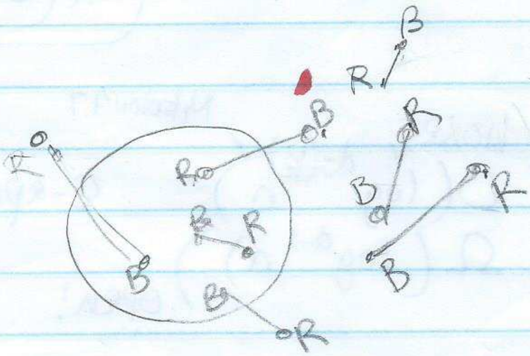
Method 1. random sampling $\Rightarrow$ discrep. $\boxed{O(\sqrt{n \log n})}$

Method 2 matching with low crossing number

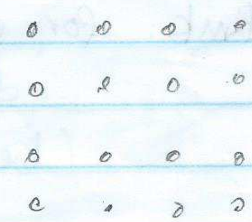
Thm (Welzl '88) Given n pts P in $\mathbb{R}^2$,

$\exists$ perfect matching M s.t. every circle (line) crosses $\le O(\sqrt{n})$ edges of M

$O(n^{1-1/d})$ in d dims



worst ex.



color each edge in M with red & blue

$\Rightarrow$ discrepancy $\boxed{O(\sqrt{n})}$

Method 3 combine! (Matoušek, Welzl - Wernisch '91)

color each edge in M with red & blue randomly

$\Rightarrow$ discrep. $O(\sqrt{\sqrt{n} \log n})$

$= \boxed{O(n^{1/4} \sqrt{\log n})}$

$\Rightarrow \exists\, O(\frac{\sqrt{\log n}}{n^{3/4}})$ - approx of size $\frac{n}{2}$

$\Rightarrow \exists\ O\left(\frac{\sqrt{\log t}}{t^{3/4}}\right)$ - approx of size t

Set $t = \frac{c}{\varepsilon^{4/3}} \log^{2/3} \frac{1}{\varepsilon}$

$\Rightarrow \exists\ \varepsilon$-approx of size $\boxed{O\left(\frac{1}{\varepsilon^{4/3}} \log^{2/3} \frac{1}{\varepsilon}\right)}$

Rmk: for ranges with (primal or dual) shattering dim d,

discrep. $\leq O(\sqrt{n^{1-\frac{1}{d}}}) = \tilde{O}(n^{\frac{1}{2}-\frac{1}{2d}})$

$\exists\ \varepsilon$-approx of size $\boxed{\tilde{O}\left(\left(\frac{1}{\varepsilon}\right)^{2-\frac{2}{d+1}}\right)}$

Rmk: for rectangles/boxes, Nikolov'17

discrep $= O(\log^{d-(1/2)} n)$ ε-approx size $O\left(\frac{1}{\varepsilon} \log^{d-\frac{1}{2}} \frac{1}{\varepsilon}\right)$

(best lower bd $\Omega(\log^{d-1} n)$) open!

efficient construction ... (Bansal, Lovett-Meka, ...)

Rmk. general sets

(Spencer, Beck-Fiala, ...)

Derandomization of ε-Approximations

polytime by method of conditional probability

faster? $O(n)$ time for const ε (Matoušek '95)

Fact: If R_1 is an ε-approx of P_1

R_2 " " " " P_2, $|P_1| = |P_2|$,

then $R_1 \cup R_2$ " " " " $P_1 \cup P_2$.

Alg'm by "merge & reduce":

to compute $\varepsilon(n)$-approx of P:

1. divide P into P_1, P_2 of size $n/2$ (arbitrarily)
2. recursively compute $\varepsilon(n/2)$-approx R_1 of P_1
   " " " R_2 of P_2
3. return $\delta(n)$-approx R of $R_1 \cup R_2$.

size $O\left(\frac{1}{\delta(n)^2}\log\frac{1}{\delta(n)}\right)$

$$\varepsilon(n) = \varepsilon(n/2) + \delta(n)$$

by slow alg'm polytime

$$T(n) = 2T\left(\frac{n}{2}\right) + O\left(\left(\frac{1}{\delta(n/2)^2}\log\frac{1}{\delta(n/2)}\right)^c\right)$$

e.g. set $\delta(n) = \delta$

$\Rightarrow$ $(\delta \log n)$-approx of size $O\left(\frac{1}{\delta^2}\log\frac{1}{\delta}\right)$

in $O\left(\left(\frac{1}{\delta}\right)^{O(1)} n\right)$ time

$\Rightarrow$ ε-approx of size $O\left(\frac{1}{\varepsilon^2}\log^{O(1)} n\right)$ ← can be reduced to $O\left(\frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon}\right)$ by slow alg'm

in $O\left(\left(\frac{1}{\varepsilon}\right)^{O(1)} n \log^{O(1)} n\right)$ time

Improvement:

Set $\delta(n) = \frac{1}{n^{\alpha}}$ for $n \geq \left(\frac{1}{\varepsilon}\right)^{1/\alpha}$.

$$\varepsilon(n) = \varepsilon(n/2) + \frac{1}{n^{\alpha}}$$

$$= \frac{1}{n^{\alpha}} + \frac{1}{(n/2)^{\alpha}} + \dots + \frac{1}{(1/\varepsilon)} = O(\varepsilon).$$

$$T(n) = \begin{cases} 2\,T\left(\frac{n}{2}\right) + O\left(\left(n^{2\alpha}\log n\right)^{c}\right) & \text{if } n \geq \left(\frac{1}{\varepsilon}\right)^{1/\alpha} \\ O\left(\left(\left(\frac{1}{\varepsilon}\right)^{1/\alpha}\right)^{c}\right) & \text{else} \end{cases} \implies T(n) = O\left(\left(\frac{1}{\varepsilon}\right)^{O(1)} n\right)$$

by setting $\alpha < \frac{1}{2c}$

$\implies$ ε-approx of size $O(n^{2\alpha}\log n)$ * in $O\left(\left(\frac{1}{\varepsilon}\right)^{O(1)} n\right)$ time

* can be reduced to $O\left(\frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon}\right)$ by slow alg'm

Rmk: in 1D, approx quantiles / median ...