CPS234 Computational Geometry **1 November 2005**

Lecture 18: Random Sampling

Lecturer: Pankaj K. Agarwal *Scribe: Amber Stillings*

18.1 Lecture Summary

This lecture will show motivating reasons for using random sampling. An algorithm for finding the median of numbers using random sampling is shown. Geometric examples of random sampling are also shown.

Random sampling methods sacrifice accuracy in favor of better space and time requirements, and methods of bounding error are described, including ϵ- nets, ϵ-approximations, and discrepancy. VC-dimension is also described and shown as a tool for bounding error.

18.2 Toy Example: Finding the Median of Numbers

$X = x_1, \ldots, x_n \subseteq \Re$
Goal: Compute the median of X
Median of medians takes linear time. Described here is another linear time algorithm, with a better constant than median of medians.

18.2.1 Random Sampling Median Algorithm

```
1) Choose $\sqrt{n}$ numbers at random
2) Sort these $\sqrt{n}$ numbers. $Y = x_{i1} < x_{i2} < \ldots < x_{ik}$ where $k = \sqrt{n}$
3) Let $X^-$ be the elements of rank $k/2 - 5$ in $Y$
   Let $X^+$ be the elements of rank $k/2 + 5$ in $Y$
4) $rk(x)$: rank of $x$ in $X$
   $Z = x | rk(X^-) < rk(x) < rk(X^+)$
5) if $|Z| \geq c \cdot \sqrt{n}$ or $((rk(X^-) < n/2 < rk(X^+)$ not true) then go to step 1
6) Compute the element of rank $n/2 - rk(X^-)$ in $Z$
```

18.3 Geometry Examples

18.3.1 Range Searching

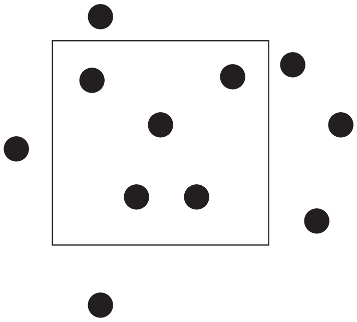


Figure 18.1: Orthogonal Range Searching: Points and Query Rectangle

Count points in query rectangle. With $O(n \log n)$ space, this can be done in $O(\log n)$ query time.

Some applications have too much data though ($O(n \log n)$ space is unreasonable). Instead, a random subset is stored. A correct answer cannot be expected. Instead, an approximation is obtained.

B: words of storage
S: Set of n points in $\Re^2$
Store a random subset $N \subseteq S$ of B points.

r: query rectangle
Compute $|r \cap N|$
Return $|r \cap N| \cdot \frac{|S|}{|N|}$

18.3.2 Query Segments

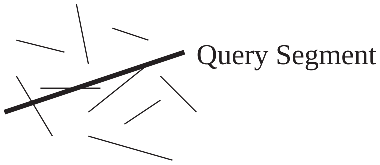


Figure 18.2: Input and Query Segments

Does the query segment intersect any of the input segments? This problem relates to motion planning in AI (Probabilistic Road Maps). A tradeoff is made. Accuracy is sacrificed in order to improve speed.

Some mistakes are made, especially if the query intersects few segments (ie, the query segment intersects only 1 segment and that segment is not in the sample being stored).

18.4 Statistical Analysis

X: finite set of n objects $\in \Re^2$
R: a set of subsets of X, $R \subseteq 2^{|X|}$

R is called ranges or hyperedges.
$\Sigma = (X, R)$: set system, hypergraph, range space

18.4.1 Ranges R_1

$R_1 = \{X \cap \gamma | \gamma \text{ is a halfplane}\}$

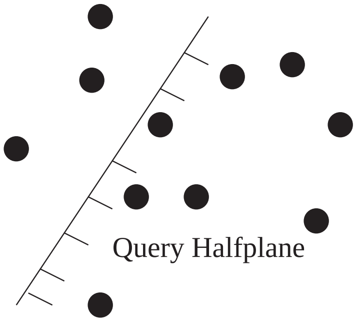


Figure 18.3: Points and Query Halfplane

$|R_1| \leq 2\binom{n}{2}$

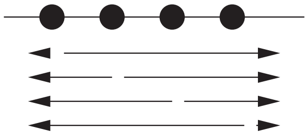


Figure 18.4: One Dimensional example of R_1, where $|R_1| = 2n$

$\mathbb{H} = (X, R_1)$

18.4.2 Ranges R_2

$R_2 = \{X \cap \rho | \rho$ is an orthogonal rectangle$\}$
$|R_2| = O(n^4), \;\; \mathbb{R} = (X, R_2)$

18.4.3 Ranges R_3

$R_3 = \{X \cap \pi | \pi$ is a convex polygon$\}$
$|R_3| \leq 2^n$

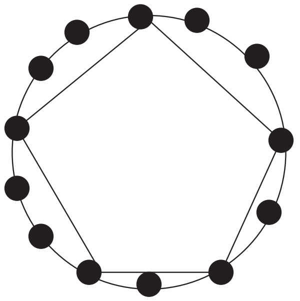


Figure 18.5: Points on a circle illustrate how $|R_3|$ is exponential

$\Pi = (X, R_3)$

18.4.4 Ranges R_4

χ: set of n lines in $\Re^2$
$R_4 = \{\{l \in \chi | l$ intersects $e\} | e$ is a segment$\}$

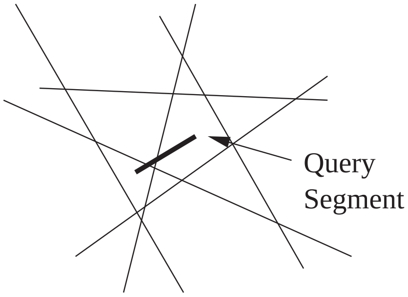


Figure 18.6: Example of a range in R_4. If the endpoints of segments are in the same face, they intersect the same lines. There are n^4 pairs of segments.

$|R_4| = O(n^4)$

18.5 ϵ-nets, ϵ-approximations, discrepancy, and VC-Dimension

Definition 1 $\Sigma = (X, R), 0 \leq \epsilon \leq 1$
$N \subseteq X$ *is an* ϵ*-net if* $\forall r \in R, |r| \geq \epsilon|X| \Rightarrow r \cap N \neq \emptyset$

ϵ-nets are like hitting sets.

Definition 2 $\Sigma = (X, R), 0 \leq \epsilon \leq 1$
$N \subseteq X$ *is an* ϵ*-approximation if* $\forall r \in R, \left| \frac{|r|}{|X|} - \frac{r \cap N|}{|N|} \right| < \epsilon$

ϵ-approximations are stronger than ϵ-nets.

Definition 3 $\Sigma = (X, R)$
$\chi : X \to \{-1, 1\}$
$disc(r) = |\sum_{x \in r} \chi(x)|$
$disc(\Sigma, \chi) = max_{r \in R} disc(r)$
$disc(\Sigma) = min_{\chi} disc(\Sigma, \chi)$

Discrepancy looks for inbalance. If the set has a small discrepancy, this means the set was balanced, and approximately half the points can be removed from the set (ie, get rid of all -1 points and use only +1 points).

How can you color the points (mark them as either -1 or 1) so that for all possible halfplanes, $\#(-1) \approx \#(+1)$?

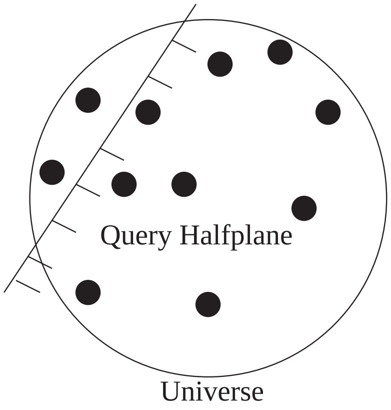


Figure 18.7: If the discrepancy is small, the number of positive and negative points in the halfplane is approximately equal.

Theorem 1 $disc(\mathbb{H}) = \Theta(n^{\frac{1}{4}})$
$disc(R_2) = \theta(\log n)$

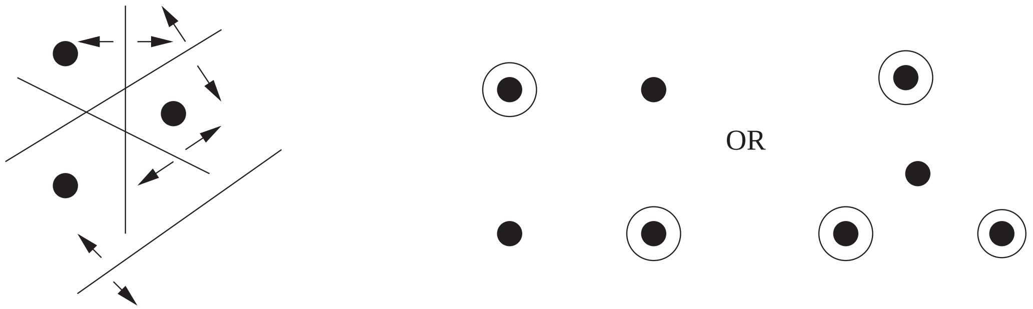
This means that there exists a coloration so that $\#(-1) = O(n^{\frac{1}{4}}) \cdot \#(+1)$ or vice versa.

After removing half the points, (say, the +1 points), you can recurse to get the total number of points desired.

18.5.1 VC-Dimension

$\Sigma = (X, R)$
$Y \subseteq X$
$R_Y = \{r \cap Y | r \in R\}$
$\Sigma_Y = (Y, R_Y)$
Y is shattered by Σ if $|\Sigma_Y| = 2^{|Y|}$

Definition 4 $VCDim(\Sigma)$*: maximum size of a subset* $Y \subseteq X$ *that can be shattered by* Σ*.*



3 points can be shattered by halfplanes

There are 2 possible configurations of 4 points. The circled points show combinations that cannot be shattered by halplanes. Therefore, 4 points cannot be shattered by halfplanes.

Figure 18.8: 3 points can be shattered by halfplanes. 4 cannot.

$VCDim(\mathbb{H}) = 3$
$VCDim(\Sigma) = \infty$ if the subsets of all sizes can be shattered.
$VCDim(\Pi) = \infty$

Claim 1 $VCDim(\Sigma) = d, \ |X| = n \ then$
$|R| \leq \sum_{i=0}^{d} \binom{n}{i} = O(n^d)$

Theorem 2 $\Sigma = (X, R), \ 0 < \epsilon < 1, \ VCDim(\Sigma) = d$
A random subset $N \subseteq X$ *of size*
$\frac{8d}{\epsilon}(\ln(\frac{1}{\epsilon}) + \ln(\frac{1}{\delta}))$
is an ϵ*-net of* Σ *with probability* $\geq 1 - \delta$*.*

Theorem 3 $\Sigma = (X, R),\ \ 0 < \epsilon < 1,\ \ VCDim(\Sigma) = d$
A random subset $N \subseteq X$ *of size*
$\frac{8d}{\epsilon^2}(\ln(\frac{1}{\epsilon}) + \ln(\frac{1}{\delta}))$
is an ϵ*-approximation of* Σ *with probability* $\geq 1 - \delta$.

There exists an ϵ-approximation of size $\Theta((\frac{1}{\epsilon})^{\frac{2d}{d+1}})$.

Next class we will talk about applications of ϵ-nets and a sketch of the proofs for these claims and theorems will be given.