Some Data Depth Problems

Questions from Statistics have inspired fundamental research in Computer Science. One example involves the notion of data depth - depth defined via a given set of data points. Let $a_1, \ldots, a_n$, $a_i \in R$ be a set of n distinct inputs. The points partition the real line into $n+1$ intervals and the depth of $x \in R$ is defined to be

$$d(x) = \min(|\{a_i : a_i \leq x\}|, |\{a_i : a_i \geq x\}|),$$

the fewest number of intervals that x meets as $|x| \to \infty$. A median is a point μ of maximal depth. It is familiar that a median has depth $\lfloor (n+1)/2 \rfloor$ and that it may be computed in linear time.

Many applications - mostly from Statistics - require an analogous notion for the depth of points $x \in R^d$. Several interesting generalizations of one-dimensional depth have been suggested. They pose a variety of challenges to computer science. On the algorithmic side, it is important to understand the complexity of computations involving depth, and to have efficient algorithms for these tasks. At the same time the various depth notions lead to combinatorial questions about arrangements of points and hyperplanes in $d-$dimensional space.

This project will address two different notions of multivariate depth, (i) *simplicial depth* and (ii) *Tukey depth*. Given a set $S = \{P_1, \ldots, P_n\}$ of n points in R^d, a $d-$simplex $\Delta P_{j_1} P_{j_2} \cdots P_{j_{d+1}}$ is the convex hull of $d+1$ distinct points from S; there are $\binom{n}{d+1}$ of them. The simplicial

$P_{j_1} P_{j_2} \cdots P_{j_{d+1}}$ is the convex hull of $d+1$ distinct points from S; there are $\binom{n}{d+1}$ of them. The simplicial depth of a point $z \in R^d$ is the number of d-simplices Δ for which $z \cap \Delta \neq \phi$. A simplicial median is a point of maximal simplicial depth. The Tukey depth of $z \in R^d$ is defined to be the minimum number of points of S in any closed halfspace containing z.

Given n points in R^2, a simplicial median can be computed in $O(n^4)$ time. Also, a lower bound of $\Omega(n \log n)$ is known for this task. A major goal of this project is to narrow the gap between these bounds. Another objective will be to shrink the analogous gap for the computational complexity of the Tukey median with n points in R^2.

Finally, some effort will be devoted towards applications of depth notions in statistics and the analysis of data.