Lower Bounds for Computing Statistical Depth

Greg Aloupis* Carmen Cortés† Francisco Gómez‡
Michael Soss* Godfried Toussaint*

February 22, 2001

Abstract

Given a finite set of points S, two measures of the depth of a query point θ with respect to S are the *Simplicial depth* of Liu and the *Halfspace depth* of Tukey (also known as *Location depth*). We show that computing these depths requires $\Omega(n \log n)$ time, which matches the upper bound complexities of the algorithms of Rousseeuw and Ruts. Our lower bound proofs may also be applied to two bivariate sign tests: that of Hodges, and that of Oja and Nyblom.

1 Introduction

The notion of depth for a point with respect to a data set has been studied extensively by statisticians and computer scientists. Applications include robust estimation, hypothesis testing, graphical display [MRR$^+$01], data description, multivariate confidence regions, p-values, quality indices, control charts [RR96], and even voting theory [RR99]. In general, the depth of a point θ quantifies the degree to which θ is centrally located in a data set. An introduction to several definitions of depth, as well as their properties, is

*School of Computer Science, McGill University. {athens,soss,godfried}@cs.mcgill.ca

†Departamento de Matematica Aplicada, Escuela Universitaria de Ingeniera Tecnica Agricola, Universidad de Sevilla. ccortes@cica.es

‡Escuela Universitaria de Informatica, Matematica Aplicada, Universidad Politecnica de Madrid. fmartin@eui.upm.es

given in [Sma90]. Below, we include descriptions of the Halfspace depth of Tukey [Tuk75] and the Simplicial depth of Liu [Liu90].

Let $S = \{s_1, \ldots, s_n\}$ be a set of data points in R^d. The *Halfspace depth* of a point $\theta \in R^d$ with respect to S is the minimum number of points in S contained in any halfspace which includes θ. The *Simplicial depth* of a point θ in R^d with respect to S is the number of simplices formed by $d+1$ elements of S that contain θ. To find the Simplicial depth of θ in R^2, we must find how many triangles formed by triples of points in S contain θ. (Halfspace and Simplicial depths are sometimes normalized by a function of n, although in this paper we will not use normalization for either.)

In section 2 we describe simplified versions of the algorithms of Rousseeuw and Ruts [RR96] for computing the Simplicial and Halfspace depth of a point in R^2. The time complexity of each algorithm is $O(n \log n)$ in the RAM model of computation. In section 3 we prove that the computation of Simplicial and Halfspace depth in R^2 requires $\Omega(n \log n)$ time in the algebraic decision tree model of computation. For a discussion on the connection between the two models of computation, we refer the reader to [PS80]. Finally in section 4 we show that the lower bounds also apply for the sign tests of Hodges [Hod55] and of Oja and Nyblom [ON89].

2 Algorithms for the Depth of a Point in R^2

The following two algorithms are simplifications of those of Rousseeuw and Ruts [RR96].

2.1 Halfspace depth calculation

Suppose we have a data set S of n points, and a point θ for which we want to compute Halfspace depth.

First, note that it suffices to consider only halfspaces determined by lines through θ. If any element in S coincides with θ we can ignore it and increment the depth value when we are finished. Sort S radially about θ and construct a directed line L through θ and some point s in S. For example, in figure 1 L is directed from θ to point 1. Let L_f be the halfline on L that extends from θ and crosses s (shown thicker in figure 1). Count the number of points on or to the left of L, excluding points on L_f. This represents the points in the closed halfspace defined by a line rotated slightly counterclockwise from

L. Rotate L counterclockwise until it encounters a new point. Notice that it is possible for L to rotate through an angle of zero. If the next encounter is on L_f, we know that the current halfspace defined by L will have one less point. Otherwise, we know that the current halfspace will gain one point. We update the minimum value found every time L changes direction (if there are no two points collinear with θ this will happen every time). The process ends when L has performed one full cycle.

In the example of figure 1, L initially contains points 6 and 1. The initial halfspace contains points 2 through 6, so the initial minimum is five. When L is rotated it first encounters point 7, which is not on L_f. Thus the new halfspace contains six points and the minimum is still five. After the next rotation, L will contain points 2,3,4 and 8. The minimum value found will be updated to four after all of these points are processed.

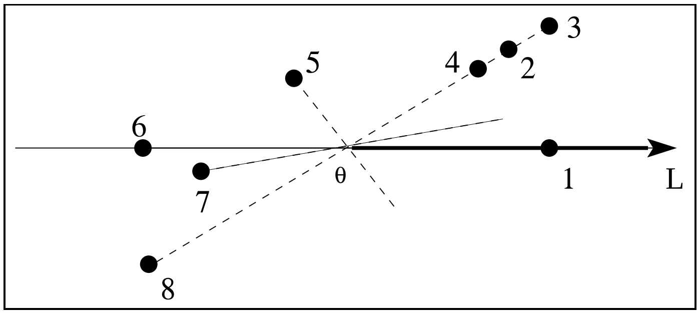

Figure 1: Sorting technique for depth computation

The overall running time is dominated by the sorting step, so it is $O(n \log n)$.

2.2 Simplicial depth calculation

To find the Simplicial depth of θ with respect to S we can compute the number of triangles formed by triples of points in S which do *not* contain θ and subtract this number from the total number of triangles. We will use the fact that the three vertices of a triangle which does not contain θ must lie in an open halfspace determined by a line through θ.

The algorithm for Simplicial depth relies on the same technique used for Halfspace depth. Consider any halfspace determined by a line through θ. Any triangle formed by points in that halfspace cannot contain θ. All that is required to compute the Simplicial depth of θ is to enumerate all such triangles and subtract them from the total number which is $\binom{n}{3}$.

Sort S radially about θ, breaking ties by placing points furthest from θ first. For each point s_i $(1 \leq i \leq n)$ that is met by the line L we compute h_i: the number of points which are between θ and s_i or strictly to the left of L. The number of triangles whose first vertex is s_i and which do not contain θ is $\binom{h_i}{2}$. By proceeding in this way we are sure that every triangle is counted once. The quantity h is updated in the same way as in the previous section.

The Simplicial depth of θ is

$$\binom{n}{3} - \sum_{i=1}^{n} \binom{h_i}{2}.$$

where $\binom{p}{q}$ is zero if $p < q$.

The complexity is identical to that of calculating Halfspace depth.

3 Lower Bounds for Computing the Depth of a Point in R^2

In this section we show that computing Simplicial and Halfspace depths requires $\Omega(n \log n)$ time.

3.1 Halfspace depth lower bound

We show that finding Halfspace depth allows us to answer the question of *Set Equality*, which has an $\Omega(n \log n)$ lower bound in the algebraic decision tree model of computation [BO83]:

- **Set Equality**: Given two sets $A = \{a_1, a_2, \ldots, a_n\}$ and $B = \{b_1, b_2, \ldots, b_n\}$, is $A = B$?

Lemma 1 *Let $S = \{s_1, s_2, \ldots, s_{2n}\}$ be a set of $2n$ points in the plane radially sorted around the point $\theta \notin S$. Then the Halfspace depth of θ is n if and only if s_i,θ,s_{n+i} are collinear, with θ between s_i and s_{n+i}, for all $1 \leq i \leq n$.*

Proof. Suppose that s_i,θ,s_{n+i} are collinear, with θ between s_i and s_{n+i}, for all $1 \leq i \leq n$. Then for any line L through θ, the points s_i and s_{n+i} either lie on opposite sides of L, or they both lie on L (see figure 2a). Since we have n such pairs of $\{s_i, s_{n+i}\}$, each closed halfspace determined by a line through θ contains at least n points of S. The minimum of n is achieved by selecting a line which does not touch any points of S.

Now suppose that s_i,θ,s_{n+i} are not collinear or that θ is not between s_i and s_{n+i} for some $1 \leq i \leq n$. Then since the angle $\angle s_i \theta s_{n+i} < 2\pi$, it is possible to draw a line L through θ such that s_i and s_{n+i} are on the same side of L. Therefore all s_j for $i \leq j \leq n+i$ are strictly on one side of L. Since there are at least $n+1$ points strictly on one side of L, there are at most $n-1$ points on the other side or on the line. Thus the depth of θ is at most $n-1$ (see figure 2b).

□

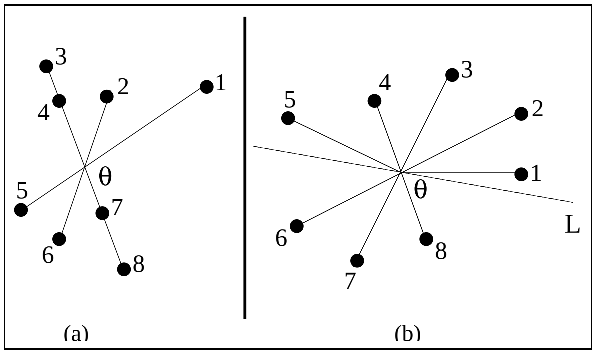

Figure 2: Halfspace depth depends on angular symmetry

Theorem 3.1 ***Planar Halfspace depth*** *requires* $\Omega(n \log n)$ *time in the worst case.*

Proof. We reduce *Set Equality* to Halfspace depth. Let $A = \{a_1, a_2, \ldots, a_n\}$ and $B = \{b_1, b_2, \ldots, b_n\}$ be two sets of real numbers. For every i ($1 \leq i \leq n$) construct the points $(a_i, 1)$ and $(-b_i, -1)$ in the plane[1]. Thus we have a set

[1] If non-degenerate points are desired, we can instead construct the points (ia_i, i) and $(-ib_i, -i)$.

S of $2n$ points, and we select $(0,0)$ as the query point θ. Considering the elements of S in radially sorted order about θ, the elements of A correspond to the points $s_1 \ldots s_n$, and the elements of B correspond to $s_{n+1} \ldots s_{2n}$. The only computation is the construction of S, which can be performed in linear time. Suppose we now find the Halfspace depth of θ. If it is n, we know that s_i,θ,s_{n+i} are collinear for all $1 \leq i \leq n$, by lemma 1. Therefore $A = B$. Again, by lemma 1, if the depth of θ is not equal to n, we cannot have n pairs of points which are reflections of each other through θ, so $A \neq B$. Therefore by finding Halfspace depth, we can answer the question of Set Equality.

□

3.2 Simplicial depth lower bound

We show that finding Simplicial depth allows us to answer the question of *Element Uniqueness*, which has an $\Omega(n \log n)$ lower bound in the algebraic decision tree model of computation [BO83]:

- **Element Uniqueness**: Given a set $A = \{a_1, a_2, \ldots, a_n\}$, is there a pair $i \neq j$ such that $a_i = a_j$?

Theorem 3.2 ***Planar Simplicial depth*** *requires* $\Omega(n \log n)$ *time in the worst case.*

Proof. We reduce *Element Uniqueness* to Simplicial depth. Let $A = \{a_1, a_2, \ldots, a_n\}$ be a set of real numbers, for $n \geq 3$. For every a_i where $1 \leq i \leq n$ construct the points $(a_i, 1)$ and $(-a_i, -1)$ which are reflections of each other through (0,0). Thus we have a set S of $2n$ points. s_i and s_{n+i} are reflections of each other through the origin, which we select as the query point θ.

Suppose s_i is a unique element in S. Then the quantity h_i, as defined in section 2.2, must equal $n-1$, since h_i includes all points $s_{i+1}, \ldots, s_{n+i-1}$ (see figure 3a). Thus if no element is duplicated in S, the Simplicial depth of θ with respect to S must be

$$D = \binom{2n}{3} - \sum_{i=1}^{2n} \binom{n-1}{2}.$$

Now suppose $s_i = s_{i+1}$ for some i. Then $h_{i+1} \leq n-2$, since h_{i+1} includes at most the points $s_{i+2}, \ldots, s_{n+i-1}$. It does not include the reflection of s_i

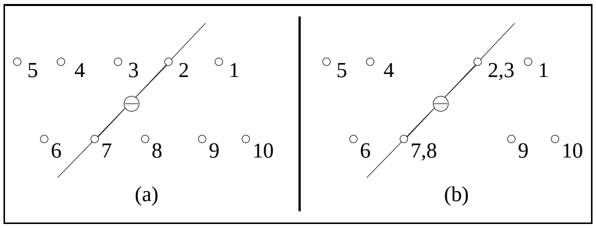

Figure 3: (a) h_2 contains s_3, s_4, s_5, s_6. — (b) h_3 contains s_4, s_5, s_6 but not s_7.

(see figure 3b). Thus if some element is duplicated in S, the Simplicial depth of θ with respect to S is strictly higher than if S has no repeated elements. Therefore by finding Simplicial depth, we can answer the question of Element Uniqueness: the elements of A are unique if and only if the depth of (0,0) with respect to S is D. The only computations in the reduction are the construction of S and the computation of D, which can be performed in linear time.

□

4 Ramifications

In this section we show that the lower bounds developed in section 3 may be applied to the bivariate sign tests of Hodges [Hod55] and of Oja and Nyblom [ON89]. Sign tests are used to determine if there is a statistically significant difference in the two distributions of n pairs of data points.

4.1 The bivariate Hodges sign test

The Hodges sign test is conducted as follows: Given a direction d and a set of n vectors formed by n pairs of points, project each vector onto d and count the number of projections which have the same direction as d. We call such vectors *positive*. The output of the test is the maximum number of positive vectors found over all directions. Rousseeuw and Ruts [RH99] mention that Halfspace depth is related to the Hodges sign test. We briefly explain how: first, shift every vector to the origin. This does not influence any directions, so the output of the test is unaffected. Now notice that for a given direction d the number of positive vectors is determined by a halfspace orthogonal to

d. It is now clear that the maximum number of positive vectors over all directions is the complement of the Halfspace depth of the origin.

Theorem 4.1 *The bivariate Hodges sign test requires* $\Omega(n \log n)$ *time in the worst case.*

Proof. The Halfspace depth of a query point θ with respect to a data set of n points may be determined by constructing n vectors (from θ to each data point) in linear time and applying the Hodges sign test at θ.

□

4.2 The bivariate Oja-Nyblom sign test

Given a data set of n points, the sign test of Oja and Nyblom involves computing the number of triples of points, each of which has the property that it falls on the same side of some line through the origin. Rousseeuw and Ruts [RR96] mention that their methods, described in section 2, may be used to compute this sign test. The relation to the problem of computing Simplicial depth is clear.

Theorem 4.2 *The Oja-Nyblom sign test requires* $\Omega(n \log n)$ *time in the worst case.*

Proof. By computing the Oja-Nyblom sign test, Simplicial depth may be determined in constant time.

5 Conclusion

We have shown that the computation of Simplicial or Halfspace depth for one point in R^2 requires $\Omega(n \log n)$ time. Thus we have matched the upper bound complexity of the algorithms of Rousseeuw and Ruts. In addition, we have shown that the computation of the bivariate sign tests of Hodges and of Oja and Nyblom require $\Omega(n \log n)$ time.

Recently we learned that a different lower bound proof for Halfspace depth has been obtained independently by Langerman and Steiger [LS00].

Acknowledgements

This research was started at the Workshop on Applications of Computational Geometry in Statistics, organized by Godfried Toussaint at the Bellairs Research Institute of McGill University in Barbados, February 1999. We thank all the participants for helpful discussions: Prosenjit Bose, Anna Bretcher, Michael Houle, Henk Meijer, Mark Overmars, Suneeta Ramaswami, Peter Rousseeuw, Toni Sellares, Diane Souvaine, Ileana Streinu and Anja Struyf.

References

[BO83] M. Ben-Or. Lower bounds for algebraic computation trees. In *Proc. 15th Ann. ACM Sympos. Theory Comput.*, pages 80–86, 1983.

[Hod55] J. Hodges. A bivariate sign test. In *Annals of Mathematical Statistics*, volume 26, pages 523–527, 1955.

[Liu90] R. Liu. On a notion of data depth based upon random simplices. *The Annals of Statistics*, 18:405–414, 1990.

[LS00] S. Langerman and W. Steiger. The complexity of hyperplane depth in the plane. In *Japan Conference on Discrete and Computational Geometry*, November 2000.

[MRR$^+$01] K. Miller, S. Ramaswami, P. Rousseeuw, T. Sellarès, D. Souvaine, I. Streinu, and A. Struyf. Fast implementation of depth contours using topological sweep. In *Proc. 12th Symposium on Discrete Algorithms (SODA)*, Washington D.C., 2001.

[ON89] H. Oja and J. Nyblom. Bivariate sign tests. *Journal of the American Statistical Association*, 84(405):249–259, 1989.

[PS80] W. Paul and J. Simon. Decision trees and random access machines. *Logic and Algorithmics, Monograph 30, L'Enseignement Mathematique*, 1980.

[RH99] P. Rousseeuw and M. Hubert. Regression depth. *Journal of the American Statistical Association*, 94:388–402, 1999.

[RR96] P. Rousseeuw and I. Ruts. Bivariate location depth. *Applied Statistics*, 45:516–526, 1996.

[RR99] P. Rousseeuw and I. Ruts. The depth function of a population distribution. *Metrika*, 49(3):213–244, 1999.

[Sma90] C. Small. A survey of multidimensional medians. *International Statistical Review*, 58:263–277, 1990.

[Tuk75] J. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, pages 523–531, Vancouver, 1975.