A Lower Bound for Computing Oja Depth

Greg Aloupis* Erin McLeish*

September 17, 2004

Abstract

Let $S = \{s_1, \ldots, s_n\}$ be a set of points in the plane. The *Oja depth* of a query point θ with respect to S is the sum of the areas of all triangles (θ, s_i, s_j). This depth may be computed in $O(n \log n)$ time in the RAM model of computation. We show that a matching lower bound holds in the algebraic decision tree model. This bound also applies to the computation of the *Oja gradient*, the *Oja sign test*, and to the problem of computing the sum of pairwise distances among points on a line.

Keywords: *algorithms, computational geometry, computational statistics, data depth, bivariate medians*

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

1 Introduction

The *depth* of a point θ with respect to a data set $S = \{s_1, \ldots, s_n\}$ in the plane is a quantitative measure of how central θ is in S. Several notions of depth exist, and typically a multivariate median definition can be formed by taking a point with minimum/maximum depth. Such notions are of great interest to the statistical community. We refer the reader to [1, 2] for recent results with a similar flavor to what is presented here, and for further introductory references to the topic of statistical depth. The main result in [1] involves lower bounds on the computation of halfspace depth [8] and simplicial depth [5]. For both depths, the given bound of $\Omega(n \log n)$ was in the algebraic decision tree model and matched known upper bounds in the RAM model. The same upper bound also applies to the computation of the *Oja depth* [6] of a point θ [7]. This depth is defined to be the sum of the areas of all triangles (θ, s_i, s_j), as shown in Figure 1.

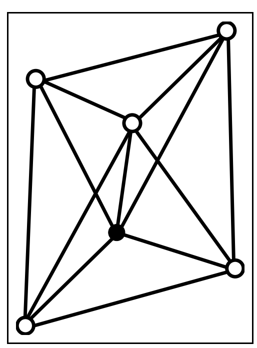

Figure 1: The Oja depth of the black point with respect to the set of white points is the sum of the areas of all triangles that include the black point as a vertex.

Here we show that the computation of Oja depth takes $\Omega(n \log n)$ time. This bound also holds for the computation of the *Oja gradient*, the *Oja sign test*, and for the problem of computing the sum of pairwise distances among points on a line.

2 The Lower Bound

We reduce the problem of *Set Equality* to the problem of computing Oja depth. It is known that determining whether two sets of real numbers are equal requires $\Omega(n \log n)$ time in the algebraic decision tree model [3]. We show that by performing some work that takes $O(n)$ time and then making a few Oja depth queries, we can answer the question of Set Equality. This implies that computing Oja depth takes $\Omega(n \log n)$ time.

Suppose that we are given two sets of real numbers, $A = \{a_1, \ldots, a_n\}$ and $B = \{b_1, \ldots, b_n\}$. We wish to know if these two sets are equal. Here, we describe a procedure that answers this question. First we assign planar coordinates to the elements of our given sets, in linear time. Every element a_i simply becomes the point $(a_i, 0)$, and the same holds for all b_i. Then we select a query point θ at $(0, 2)$. The area of any triangle formed by θ and two points taken from A and/or B is equal to the distance between those two points. Thus, in our construction, the Oja depth of θ with respect to some set S is the sum of pairwise distances among elements of S. Let d_S denote this depth (or sum of distances).

Theorem 2.1 *Computing the Oja depth of a point with respect to a set of n points in the plane requires $\Omega(n \log n)$ time in the algebraic decision tree model.*

Proof:

Suppose that we are given two sets of real numbers A and B, each of size n. We first construct corresponding planar point sets as described above, and compute the Oja depths d_A, d_B and $d_{A\cup B}$. The claim is that sets A and B are equal if and only if $2(d_A + d_B) = d_{A\cup B}$. This immediately implies that computing Oja depth must take $\Omega(n \log n)$ time.

Let $C = A \cup B$, and reorder the elements in C so that they are in increasing order: $c_1 \leq c_2 \leq \ldots \leq c_{2n}$. We can rewrite d_C as a sum of the length of the intervals $|c_{i+1} - c_i|$ by noticing that the interval length is added to the sum d_C exactly $i(2n-i)$ times. So $d_C = \sum_{i=1}^{2n-1} i(2n-i)|c_{i+1} - c_i|$. Since i is the number of points in C which are smaller than or equal to c_i, we can write $i = p_i + q_i$ where $p_i = |a \in A, a \leq c_i|$ and $q_i = |b \in B, b \leq c_i|$. Thus $d_C = \sum_{i=1}^{2n-1}(p_i + q_i)(2n - p_i - q_i)|c_{i+1} - c_i|$. Similarly, we can use these intervals to rewrite the sums d_A and d_B. In the sum d_A, the value $|c_{i+1} - c_i|$ will be added exactly $p_i(n-p_i)$ times. So $d_A = \sum_{i=1}^{2n-1} p_i(n-p_i)|c_{i+1} - c_i|$ and $d_B = \sum_{i=1}^{2n-1} q_i(n-q_i)|c_{i+1} - c_i|$. Now we are ready to compare $2(d_A + d_B)$ to d_C. From the above expressions it is easy to check that

$$d_C - 2(d_A + d_B) = \sum_{i=1}^{2n-1} |c_{i+1} - c_i|(p_i - q_i)^2$$

Now assume that the sets A and B are equal. Then clearly $p_i = q_i$ for all i, and thus $2(d_A + d_B) = d_C$. On the other hand, assume the sets A and B differ on at least one point. Then there exists some i for which $p_i \neq q_i$. We can see this by selecting the first i in order for which $a_i \neq b_i$. For this i we will have that $p_i \neq q_i$. Therefore, if $A \neq B$ we have that

$$d_C - 2(d_A + d_B) = \sum_{i=1}^{2n-1} |c_{i+1} - c_i|(p_i - q_i)^2 > 0$$

Thus, using a few simple steps that take $O(n)$ time, and computing the Oja depths for sets A, B and $A \cup B$ we can answer the question of Set Equality. This implies that the computation of Oja depth requires $\Omega(n \log n)$ time.

■

3 Remarks

Given two points in the plane, define a vector that has magnitude equal to the distance between the two points, and that has direction orthogonal to the line through the points and pointed away from the origin. Given a set S and a query point θ, taken to be the origin for simplicity, the *Oja gradient* at θ is the sum of all vectors defined by pairs of points in S (see Figure 2). This gradient may be computed in $O(n \log n)$ time [7]. It is used for the computation of the Oja median, which is the point in the plane that has highest Oja depth (see [2]).

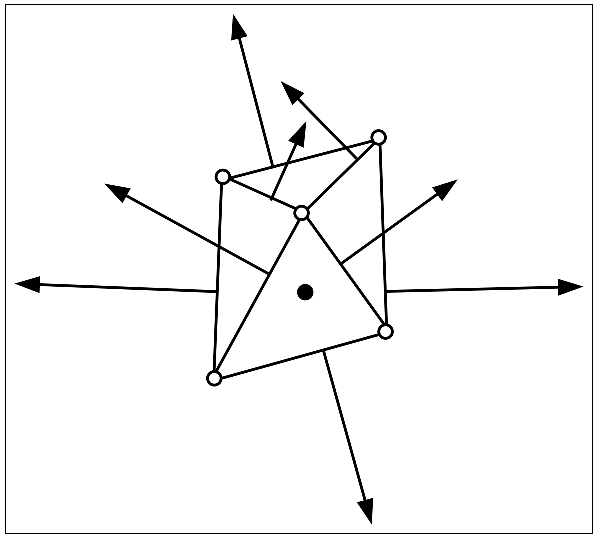

Figure 2: The Oja gradient of the black vertex with respect to the set of white vertices is a sum of vectors, each of which is orthogonal to a segment between two white vertices. Not all segments and vectors are illustrated.

Suppose that we are given a planar point set that is contained on a line ℓ. The Oja gradient at any point at unit distance from ℓ will have the same value as the Oja depth of a point that is two units of distance from ℓ. This implies that computing the Oja gradient takes $\Omega(n \log n)$ time.

Finally, Brown and Hettmansperger show that the Oja gradient can be used as a bivariate sign test [4]. Thus our lower bound holds in this case too.

References

[1] G. Aloupis, C. Cortes, F. Gomez, M. Soss, and G. Toussaint. Lower bounds for computing statistical depth. *Computational Statistics and Data Analysis*, 40:223–229, 2002.

[2] G. Aloupis, S. Langerman, M. Soss, and G. Toussaint. Algorithms for bivariate medians and a Fermat-Torricelli problem for lines. *Computational Geometry*, 26, 2003.

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

[4] B.M. Brown and T.P. Hettmansperger. An afine invariant bivariate version of the sign test. *J.R. Statist. Soc. B*, 51(1):117–125, 1989.

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

[6] H. Oja. Descriptive statistics for multivariate distributions. *Statistics and Probability Letters*, 1:327–332, 1983.

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

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