DIMACS Technical Report 2003-28
September 2003

Simplicial Depth: An Improved Definition, Analysis, and Efficiency for the Finite Sample Case

by

Michal A. Burr[1]
Dept. of Computer Science
Tufts University
Medford, Massachusetts 02155

Eynat Rafalin[1]
Dept. of Computer Science
Tufts University
Medford, Massachusetts 02155

Diane L. Souvaine[1, 2]
Dept. of Computer Science
Tufts University
Medford, Massachusetts 02155

[1]Research supported in part by NSF grant #EIA-99-96237, by NSF grant STC–91–19999 and the NJ Commission support.
The report contains material that was reported in the DIMACS Workshop on Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications and new results due in part to the participation of the authors in the workshop.
[2]Permanent Member

ABSTRACT

As proposed by Liu [12] the *simplicial depth* of a point x with respect to a probability distribution F on $\mathbb{R}^d$ is the probability that x belongs to a random simplex in $\mathbb{R}^d$. The simplicial depth of x with respect to a data set $\mathcal{S}$ in $\mathbb{R}^d$ is the fraction of the closed simplices given by $d+1$ of the data points containing the point x. We propose an alternative definition for simplicial depth which continues to remain valid over a continuous probability field, but also fixes some of the problems in the finite sample case, including those discussed by Zuo and Serfling [18]. Additionally, we discuss the effect of the revised definition on the efficiency of previously developed algorithms and prove tight bounds on the value of the simplicial depth based on the half-space depth.

1 Introduction

Statistical data depth functions play an increasing role in multivariate non-parametric analysis. A data depth measures how deep (or central) a given point $x \in \mathbb{R}^d$ is relative to F, a probability distribution in $\mathbb{R}^d$ or relative to a given data cloud. Some examples of data depth functions are *Halfspace Depth* [10, 17], *Mahalanobis Depth* [14], *Majority Depth* [16], *Simplicial Depth* [12] and the *Convex Hull Peeling Depth* [4, 7].

Most depth functions are defined in respect to a probability distribution F, considering $\{X_1, .., X_n\}$ random observations from F. The *sample version* of the depth function is obtained by replacing F by F_n, the empirical distribution of the sample $\{X_1, .., X_n\}$. This paper discusses the finite sample version of *simplicial depth*, although some references will be made to the continuous case.

To distinguish between the depth of points $\{X_i\}$ from the original data set and the depth of any other point, points which are not part of the data set are referred to as *positions*. The term *facet* is used to define the facets of a specific simplex defined by $d+1$ data points. The facets subdivide $\mathbb{R}^d$ into regions. A *cell* is the set of all positions connected by a path which does not intersect a facet.

1.1 Desirable properties of depth functions

Several properties of depth functions were introduced by Liu [12]. Recently Serfling and Zuo [18] formulated a general definition of desirable properties of depth functions, based on Liu's work, and evaluated several depth functions according to these properties. The desirable properties [18] are:

P1. Affine invariance: The depth of a point x should not depend on the underlying coordinate system, or, in particular, on the scales of the underlying measurements.

P2. Maximality at center: For a distribution having a uniquely defined center (e.g. a point of symmetry with respect to some notion of symmetry), the depth function should attain maximum value at the center.

P3. Monotonicity Relative to Deepest Point: As a point x moves away from the 'deepest point' (the point at which the depth function attains maximum value) along any fixed ray through the center, the depth at x should decrease monotonically.

P4. Vanishing at Infinity: The depth of a point x should approach zero as $||x||$ approaches ∞.

For applications of data mining and classification of large data sets, consistency under dimensions change would be another desirable property. We propose:

P5. Invariance under dimensions change: The relative depth of any two points should not depend on the dimension in which the depth was computed.

Some depth based methodologies [13, 4] are based on the *rank method*, ordering the data points according to their depth, from the center outward. In this approach the relative order of the data points is of importance and is used to extract statistical information about the data set. Note that if the maximality at center property does not hold for the data points this ordering is compromised.

1.2 Simplicial Depth Background

Simplicial depth was introduced by Liu [12] as a depth function that is robust and affine invariant. Serfling and Zuo found that the function does not behave well in some finite sample cases (under the desirable properties described above) and may be unattractive for some forms of statistical analysis. Our revised definition removes some of their concerns and alleviates other problems as well.

Definition 1. Simplicial depth (Liu [12]): Given a probability distribution F in $\mathbb{R}^d$, the *simplicial depth* of x is the probability that x belongs to a random closed simplex in $\mathbb{R}^d$:

$$SD_{Liu}(\mathcal{F}; x) = P_F(x \in S[X_1, \ldots X_{d+1}])$$

where $S[X_1, \ldots X_{d+1}]$ is a closed simplex formed by $d+1$ random observations from F[1].

Definition 2. Simplicial depth for the sample version (Liu [12]): The *simplicial depth* of a point x with respect to a data set $\mathcal{S} = \{X_1, \ldots, X_n\}$ is the fraction of the closed simplices formed by $d+1$ points of $\mathcal{S}$ containing x, where I is the indicator function:

$$SD_{Liu}(\mathcal{S}; x) = \binom{n}{d+1}^{-1} \sum I_{(x \in S[X_{i_1}, \ldots X_{i_{d+1}}])}.$$

The lower bound for computing the simplicial depth of a point or a position x relative to S in $\mathbb{R}^2$ is $\Omega(n \log n)$ [9, 1]. Several asymptotically matching $O(n \log n)$ algorithms exist ([15, 9, 11], see Section 3). For $\mathbb{R}^3$, Gil *et al.* [9] gave an $O(n^2)$ algorithm for computing the simplicial depth of x relative to S that was recently improved by Cheng and Ouyang [6]. Cheng and Ouyang also give a generalization of this algorithm to $\mathbb{R}^4$, with a time complexity of $O(n^4)$. Rousseeuw and Ruts proposed an $O(n^2)$ algorithm for $\mathbb{R}^3$ [15], but some missing details may not be resolvable [6]. For space higher than 4 dimensional there are no known algorithms faster than the straight-forward $O(n^{d+1})$ method (generate all simplices and count the number of containments). The depth of all n data points in $\mathbb{R}^2$ can be computed in $O(n^2)$ using the duality transform [9, 11].

The *simplicial median* is the point with the highest simplicial depth. In $\mathbb{R}^d$, Boros and Furedi [5] proved that the median simplicial depth is $(1/2^d)\binom{n}{d+1} + O(n^d)$. For $\mathbb{R}^2$, 1/4 is the best constant possible (the median simplicial depth is at most a 1/4 of all possible triangles from S). Aloupis *et al.* [2] showed that in 2 dimensions it can be computed in $O(n^4)$ (see Section 3). Boros and Furedi also established the existence of a position in $\mathbb{R}^2$ that is contained in exactly $2/9^{th}$ of the triangles, where $2/9^{th}$ is the best constant possible. Barany [3] generalized this by showing the existence of a position in $\mathbb{R}^d$ covered by a constant fraction of $d+1$ simplices $SD(x) = (1/[d+1^{d+1}])\binom{n}{d+1}$ [3].

1.3 Problems in the Finite Sample Case of Simplicial Depth

Several problems arise in the finite sample case of simplicial depth under Liu's definition, as detailed below. Section 3 will describe how the revised definition alleviates these problems.

[1] $S[X_1, \ldots X_{d+1}]$ represents the convex hull of the $d+1$ points. If the point set $X_1, \ldots X_{d+1}$ is not affinely independent then $S[X_1, \ldots X_{d+1}]$ is not a simplex in $\mathbb{R}^d$ but rather a convex object contained within a k-flat for $k < d$.

Figure 1: The Zuo-Serfling Counterexamples [18] (Note that Zuo and Serfling divide the number of simplicies enclosing the query point by a factor of n^3, while we use a factor of $\binom{n}{3}$, yielding different simplicial depth values): *Counterexample 1:* Let $d = 1$ and $S_1 = \{-2, -1, 0, 1, 2\}$. The center is clearly $x = 0$. $SD_{Liu}(p_1 = 1/2) = \frac{6}{10}$ while $SD_{Liu}(x_1 = 1) = \frac{7}{10}$, violating the monotonicity property P3. According to the revised definition $SD_{BRS}(p_1 = 1/2)$ is unchanged and $SD_{BRS}(x_1 = 1) = \frac{5}{10}$. *Counterexample 2:* Let $d = 2$ and $S_2 = \{(\pm 1, 0), (\pm 2, 0), (0, \pm 1)\}$. S_2 is centrally symmetric about $(0, 0)$. $SD_{Liu}(p_2 = (\frac{1}{2}, 0)) = \frac{10}{20}$ while $SD_{Liu}(x_2 = (1, 0)) = \frac{12}{20}$, again violating the monotonicity property P3. Under the revised definition $SD_{BRS}(p_2) = \frac{12}{20}$ and $SD_{BRS}(x_2) = \frac{9}{20}$, Counterexample 3 compares the depth of degenerate, multiple, points and is not described here because the data points are currently only partially treated under the revised definition (Section 6.2).

Maximality and Monotonicity: The Simplicial Depth function is a statistical depth function, in the sense of Serfling and Zuo's definition [18], for a continuous angularly symmetric distribution. Zuo and Serfling show, however, that the simplicial depth function for the finite sample case fails to satisfy the maximality (P2) and monotonicity (P3) properties using several counterexamples, two of them presented in a slightly modified form in Figure 1. The revised definition, as described below resolves the problems raised by these counterexamples. Nonetheless, as shown in Section 6, the maximality and monotonicity properties still do not hold.

Positions on Facets: Depth of positions on facets causes discontinuities in the depth function. The depth of all positions on the boundary of a cell is at least the depth of a position on the interior of the cell. In most cases the depth values on the boundaries can be higher than the depth in each of the adjacent cells (see e.g. Figure 3(a)).

2 Revised Definition

Definition 3. Revised Simplicial Depth: Given a data set $\mathcal{S} = \{X_1, \ldots, X_n\}$ in $\mathbb{R}^d$, the simplicial depth of a point x is the average of the fraction of closed simplicies containing x and the fraction of open simplicies containing x:

$$SD_{BRS}(\mathcal{S}; x) = \frac{1}{2}\binom{n}{d+1}^{-1}\left(\sum I_{(x \in S[X_{i_1}, \ldots X_{i_{d+1}}])} + I_{(x \in int(S[X_{i_1}, \ldots X_{i_{d+1}}]))}\right)$$

where *int* refers to the open relative interior[2] of $S[X_{i_1}, \ldots, X_{i_{d+1}}]$. Equivalently, this could be formulated as: $SD_{BRS}(\mathcal{S}; x) = \rho(\mathcal{S}, x) + \frac{1}{2}\sigma(\mathcal{S}, x)$ where $\rho(\mathcal{S}, x)$ is the number of simplicies with data points as vertices which contain x in their open interior, and $\sigma(\mathcal{S}, x)$ is the number of simplicies with data points as vertices which contain x in their boundary.

[2] See Edelsbrunner [8], page 401.

2.1 Properties of the Revised Definition

It can easily be seen that for continuous distributions and for positions lying in the interior of cells, the revised definition reduces to Liu's original definition. Significantly, the revised definition corrects irregularity at boundaries of simplicies presented in Section 1.3 and Figure 3(a), making the depth of a point on the boundary between two cells the average of the depth of the two cells (Proposition 1). The Zuo-Serfling counterexamples [18] are also all solved by the revised definition (see Figure 1).

Lemma 1. The simplicial depth of any two positions in the same cell is equal.

Proof. Let A be a cell defined by a d-dimensional data set $\mathcal{S} = \{X_1, \ldots, X_n\}$, and let x_{A_1} and x_{A_2} be two positions in A. As cells are convex, the segment with endpoints x_{A_1} and x_{A_2} lies entirely within A. Additionally, as x_{A_1} and x_{A_2} are two positions in the same cell, no facet intersects the segment between them. Assume that x_{A_1} does not have the same simplicial depth as x_{A_2}; w.l.o.g. let $SD(\mathcal{S}; x_1) > SD(\mathcal{S}; x_2)$. Consequently there exists a simplex which contains the point x_{A_1} but not x_{A_2}. By the generalization of the Jordan Curve theorem, the segment with endpoints of x_{A_1} and x_{A_2} must then intersect the boundary of this simplex, producing a contradiction as no facets between data points can intersect the segment with endpoints x_{A_1} and x_{A_2}. $\square$

Proposition 1. The simplicial depth of a position on a facet between two cells is equal to the average of the depths of a position in the two adjacent cells, assuming that only d points lie on the hyperplane defined by the facet.

Proof. Given a data set as defined in the proposition, consider two adjacent cells, A and B. Let x_A be a position in cell A, x_B be a position in cell B, and θ be a position which lies on the shared facet of cells A and B. The simplicial depth of these three positions is $SD(\mathcal{S}; x_A) = \rho(\mathcal{S}, x_A) + \frac{1}{2}\sigma(\mathcal{S}, x_A)$, $SD(\mathcal{S}; x_B) = \rho(\mathcal{S}, x_B) + \frac{1}{2}\sigma(\mathcal{S}, x_B)$, and $SD(\mathcal{S}; \theta) = \rho(\mathcal{S}, \theta) + \frac{1}{2}\sigma(\mathcal{S}, \theta)$. As both x_A and x_B are in the interior of cells A and B respectively, no simplicies can contain x_A or x_B in their boundaries, so $\sigma(\mathcal{S}, x_A) = \sigma(\mathcal{S}, x_B) = 0$, which results in the following depths for the three data positions: $SD(\mathcal{S}; x_A) = \rho(\mathcal{S}, x_A)$, $SD(\mathcal{S}; x_B) = \rho(\mathcal{S}, x_B)$ and $SD(\mathcal{S}; \theta) = \rho(\mathcal{S}, \theta) + \frac{1}{2}\sigma(\mathcal{S}, \theta)$

Consider first a simplex R which contains θ in its interior and thus contributes to $\rho(\mathcal{S}, \theta)$. As θ is contained in the interior of R, $\exists \epsilon > 0$ such that the d-dimensional disk $d_{(\theta,\epsilon)}$ centered at θ with radius ϵ is contained entirely within R. As θ lies on the boundary of cells A and B, $d_{(\theta,\epsilon)}$ must contain some points of A and some points of B. By Lemma 1, this implies that the simplex R contains cells A and B. Consider now a simplex T which contains θ in its boundary and thus contributes to $\sigma(\mathcal{S}, \theta)$. As the facet which includes θ is a boundary of T, the interior of the simplex lies to one side of this facet and the exterior to the other. Moreover, as only one side of the facet is in the interior of T, exactly one of A or B, lies inside T. Now let $\rho_{x_A}(\mathcal{S}, \theta)$ be the number of simplicies which contain both θ and x_A in their interiors and let $\sigma_{x_A}(\mathcal{S}, \theta)$ be the number of simplicies which contain θ in their boundary and x_A in their interior. Define $\rho_{x_B}(\mathcal{S}, \theta)$ and $\sigma_{x_B}(\mathcal{S}, \theta)$ similarly. By the above argument, $\rho_{x_A}(\mathcal{S}, \theta) = \rho_{x_B}(\mathcal{S}, \theta) = \rho(\mathcal{S}, \theta)$. Finally, consider a simplex, U, which contains cell A, then

the point θ is in the closed simplex U, as $\forall\epsilon > 0$, $d_{(\theta,\epsilon)}$ contains some points of cell A. Thus $SD(\mathcal{S}; x_A) = \rho_{x_A}(\mathcal{S}, \theta) + \sigma_{x_A}(\mathcal{S}, \theta)$; similarly, $SD(\mathcal{S}; x_B) = \rho_{x_B}(\mathcal{S}, \theta) + \sigma_{x_B}(\mathcal{S}, \theta)$. The depth of θ is therefore $SD(\mathcal{S}; \theta) = \rho(\mathcal{S}, \theta) + \frac{1}{2}\sigma(\mathcal{S}, \theta) = \frac{1}{2}\left(2\rho(\mathcal{S}, \theta) + \sigma(\mathcal{S}, \theta)\right)$
$= \frac{1}{2}\left(\rho_{x_A}(\mathcal{S}, \theta) + \rho_{x_B}(\mathcal{S}, \theta) + \sigma_{x_A}(\mathcal{S}, \theta) + \sigma_{x_B}(\mathcal{S}, \theta)\right) = \frac{1}{2}\left(SD(\mathcal{S}; x_A) + SD(\mathcal{S}; x_B)\right)$ □

Corollary 1. For a data set $\mathcal{S} = \{X_1, \ldots X_n\}$ in general position, the value of the median is attained in the interior of a cell or at a data point.

Proof. Follows from Proposition 1. □

Definition 4. Opposite Cells: Two cells whose boundaries both contain a position θ lying on the intersection of two or more hyperplanes induced by the data set are *opposite cells* if and only if the two cells lie on opposite sides of every facet that contains θ. It can be shown that every cell has a unique opposite.

Proposition 2. The depth of the position at the intersection between two or more facets is equal to the average of the depths of two opposite cells of the intersection point, assuming only d points lie on any hyperplane defined by the facets.

Proof. Following the proof for Proposition 1 where one cell is labeled A and its opposite cell B. □

Proposition 3. For a data set $\mathcal{S} = \{X_1, \ldots, X_n\}$ in $\mathbb{R}^d$, in general position, the ordering of data points by their simplicial depth due to Liu's definition is unchanged by the revised definition.

Proof. Given $\rho(\mathcal{S}, x)$ and $\sigma(\mathcal{S}, x)$ as per Definition 3 $SD_{Liu}(\mathcal{S}; x) = \rho(\mathcal{S}, x) + \sigma(\mathcal{S}, x)$ while $SD_{BRS}(\mathcal{S}; x) = \rho(\mathcal{S}, x) + \frac{1}{2}\sigma(\mathcal{S}, x)$. Based on general position, $\sigma(\mathcal{S}, X_i) = \binom{n-1}{2}$ for any data point X_i. Therefore, if $SD_{Liu}(\mathcal{S}; X_i) \leq SD_{Liu}(\mathcal{S}; X_j)$ then $SD_{BRS}(\mathcal{S}; X_i) \leq SD_{BRS}(\mathcal{S}; X_j)$. □

2.2 Invariance Under Dimensions Change - Comparing $\mathbb{R}^2$ and $\mathbb{R}$

Consider the case where a data set $\mathcal{S}_a = \{X_1, \ldots, X_n\}$ $(n \geq 3)$ consisting of a set of collinear points is analyzed as an $\mathbb{R}^2$ data set instead of as an $\mathbb{R}^1$ data set. Assume w.l.o.g. that these n points lie on the positive x-axis (see Figure 2(a)). Table 1 compares the depth assigned by Liu's definition and by the revised definition for $\mathbb{R}$ and $\mathbb{R}^2$, where m represents the number of data points which lie to the left of the position or data point under consideration.

In $\mathbb{R}^2$, for a degenerate simplex BCD, with C between B and D (see Figure 2(b)), the revised definition assumes that a position A between B and C lies within the open simplex BD and thus achieves depth of 1. Both points B and D achieve a depth of 1/2 as degenerate vertices of the simplex, whereas C earns a depth of 1 as well, as it lies within the open (degenerate) simplex BD. Liu's definition assigns a depth of 1 to all points and positions on the interval $[B, D]$.

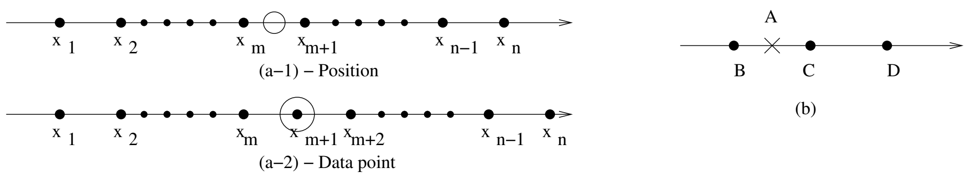


Figure 2: Data set S_a

Data Points	Dimension	Definition
$\binom{n}{2}^{-1}(m(n-m-1)+(n-1))$	$\mathbb{R}$	Liu
$\binom{n}{2}^{-1}\left(m(n-m-1)+\frac{1}{2}(n-1)\right)$	$\mathbb{R}$	Revised Definition
$\binom{n}{3}^{-1}\left((n-m-1)\binom{m}{2}+m\binom{n-m-1}{2}+\binom{n-1}{2}\right)$	$\mathbb{R}^2$	Liu
$\binom{n}{3}^{-1}\left((n-m-1)\binom{m}{2}+m\binom{n-m-1}{2}+m(n-m-1)+\frac{1}{2}\left(\binom{m}{2}+\binom{n-m-1}{2}\right)\right)$	$\mathbb{R}^2$	Revised Definition
Position	Dimension	Definition
$\binom{n}{2}^{-1}(m(n-m))$	$\mathbb{R}$	Both Definitions
$\binom{n}{3}^{-1}\left((n-m)\binom{m}{2}+m\binom{n-m}{2}\right)$	$\mathbb{R}^2$	Both Definitions

Table 1: Depth of data points and positions for set $\mathcal{S}_a$

For both definitions, the ratio of the depths of a *position* when $\mathcal{S}_a$ is analyzed first as an $\mathbb{R}^1$ data set and second as an $\mathbb{R}^2$ data set is: $\frac{\binom{n}{2}^{-1}(m(n-m))}{\binom{n}{3}^{-1}\left((n-m)\binom{m}{2}+m\binom{n-m}{2}\right)} = \frac{2}{3}$.
The same ratio for *points* rather than positions, by Liu's definition is *not* identical to $\frac{2}{3}$:

$$\frac{\binom{n}{2}^{-1}(m(n-m-1)+(n-1))}{\binom{n}{3}^{-1}\left((n-m-1)\binom{m}{2}+m\binom{n-m-1}{2}+\binom{n-1}{2}\right)} \neq \frac{2}{3}.$$

The ratio for data points under the *revised* definition, however, is uniformly $\frac{2}{3}$:

$$\frac{\binom{n}{2}^{-1}\left(m(n-m-1)+\frac{1}{2}(n-1)\right)}{\binom{n}{3}^{-1}\left((n-m-1)\binom{m}{2}+m\binom{n-m-1}{2}+m(n-m-1)+\frac{1}{2}\left(\binom{m}{2}+\binom{n-m-1}{2}\right)\right)} = \frac{2}{3}.$$

As both ratios for the revised definition are equal to $\frac{2}{3}$, this means that the only difference between a data set analyzed by the $\mathbb{R}^1$ depth measure and the $\mathbb{R}^2$ depth measure is a constant multiple. This property is desirable as discussed in Section 1, property P5. Although it holds for $\mathbb{R}^2/\mathbb{R}^1$, additional work is needed to generalize it for higher dimensions, as discussed in Section 6.2.

3 Existing Algorithms and the Effect of the Revised Definition

3.1 The Depth of a Point or Position in $\mathbb{R}^2$

The asymptotically optimal algorithm for calculating simplicial depth of a point or position in $\mathbb{R}^2$, by Rousseeuw and Ruts [15] can be modified slightly to compute the revised simplicial depth. The original algorithm computes the number of simplicies which contain the query point by counting the number of simplicies which do not contain the query point. For each $i = 1, \ldots, n$, let α_i be the angle of the vector $X_i - x$, and assume $0 = \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_n \leq 2\pi$. Let h_i be the largest integers such that $\alpha_i \leq \alpha_{i+1} \leq \cdots \leq \alpha_{i+h_i} < \alpha_i + \pi$ (in radial order).

The authors base their algorithm on the fact that a data triangle $\triangle X_i X_j X_k$ excludes x if and only if there exists an angle smaller than π which contains all three angle $\alpha_i, \alpha_j, \alpha_k$, i.e. one of the indices i, j, k comes before the other two, and these two indices are among the corresponding h_i indices. Therefore, the number of data triangles containing x is

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

Instead of computing h_i the authors compute $F(i) = h_i + i = \#\{j; 0 \leq \alpha_j < \alpha_i + \pi\}$. $F(i)$ can be computed by an updating mechanism from the ranks of the antipodal angles. Another step is needed to account for the data points that coincide with θ. The algorithm is therefore composed of two sorting steps: one that computes and orders the α_i's and another that orders the α_i's and their antipodal points in the same array. The latter array is used to count the $F(n)$'s (and h_i) and calculate the sum (1). The total time complexity is $O(n \log n)$.

The revised algorithm must compute the number of simplicies which do not contain the query point θ, but subtract half of the number of simplicies which include θ in the boundary. For each $i = 1, \ldots, n$, let h_i and k_i be the largest integers such that : $\alpha_i \leq \alpha_{i+1} \leq \cdots \leq \alpha_{i+h_i} < \alpha_{i+h_i+1} = \cdots = \alpha_{i+k_i} = \alpha_i + \pi$. Where $k_i - h_i$ is the number of antipodal points to X_i. The revised simplicial depth is:

$$\binom{n}{3} - \sum_{i=1}^{n} \left(\binom{h_i}{2} + \frac{1}{2}(k_i - h_i)h_i \right) \tag{2}$$

k_i can be computed using the same updating mechanism used to compute h_i (via $F(i)$). Therefore computation of this new summation can be done with the same complexity as described above.

3.2 The Depth of a Point or Position in $\mathbb{R}^3$

An $O(n^2)$ time algorithm for $\mathbb{R}^3$ was described by Cheng and Ouyang [6], based on Gil, Steiger and Wigderson's algorithm [9]. The authors consider the intersection points $\{\theta_i\}$ of the ray from the query point x to each X_i with the unit sphere centered around x. Clearly, containment of x in the simplex $\triangle X_i X_j X_k X_l$ is identical to containment in the simplex $\triangle \theta_i \theta_j \theta_k \theta_l$. In addition, x is contained in simplex $\triangle \theta_i \theta_j \theta_k \theta_l$ if and only if the spherical triangle $\triangle_s \theta_i \theta_j \theta_k$ contains the antipodal point θ'_l of θ_l [3]. If a plane through x that separates θ_i from $\triangle_s \theta_j \theta_k \theta_l$ is translated away from the origin, the simplex $\triangle \theta_i \theta_j \theta_k \theta_l$ contains the origin if and only if the triangle formed by the radial projection of $\triangle_s \theta_j \theta_k \theta_l$ on the plane contains

[3]A spherical triangle $\triangle_s \theta_i \theta_j \theta_k$ is the area on the surface of the unit sphere bounded by the short arcs of the great circles passing through any pair of point $\{\theta_i, \theta_j, \theta_k\}$.

the radial projection of θ_i'. Hence the 3 dimensional problem can be reduced to 2 dimensional containment problems.

The algorithm begins by projecting the data set $\mathcal{S}$ onto a plane orthogonal at x to the *axis of rotation*, a ray connecting x to a point in general position with $\mathcal{S} \cup x$, by adding all antipodal points, and by picking all *dividing planes* that contain the axis of rotation and the line that separates a projected antipodal point from its predecessor in the radial order. The problem focuses on computing, for every dividing plane, the sum of the simplicial depths of every antipodal point in the projection of points from the upper hemisphere of $\mathcal{S}$. As the algorithm moves from one dividing plane to the other, one antipodal point θ_i' moves to the upper hemisphere (new antipodal point) and some points $\theta_{j+1}, \theta_{j+2}, \ldots, \theta_{j+k}$ move to the upper hemisphere (new primary points). The following is counted: (1) For each old antipodal point, the number of triangles containing it where one or more vertices are new primary points and the rest of the vertices are old primary points. (2) For the new antipodal point, the number of triangles containing it with three primary points as vertices. For every dividing plane, the algorithm uses the dual to count containment with total time complexity $O(n^2)$. To move from one dividing plane to the next, points from the 2-dimensional arrangement that move to the lower hemisphere are removed and those that move to the upper hemisphere are added. Since each point appears and disappears once and the update time for a point is linear, this step takes $O(n^2)$ time. The total time complexity is $O(n^3)$. Cheng and Ouyang generalize the algorithm to $\mathbb{R}^4$, with $O(n^4)$ time.

This algorithm also remains valid under the revised definition, with adjustments for counting the number of points that lie on a facet. If the new antipodal point lies on the boundary of a simplex of three primary points, then x lies on a facet (easily determined, with minor changes). In addition, if three points determine a great circle and the antipodal point lies inside the triangle defined by these points, then x also lies on the boundary. To assist with this we check, for every primary point that is added to the upper hemisphere or about to be removed from it, whether it lies on the hemisphere's boundary.

3.3 The Simplicial Median in $\mathbb{R}^2$

According to Liu's definition the depth function achieves local maximum on intersection points of segments between data points. To find the simplicial median only these $O(n^4)$ intersection points have to be considered. Aloupis *et al.* [2] describe an $O(n^4)$ time and space or $O(n^4 \log n)$ time and $O(n^2)$ space algorithm for computing the simplicial median by finding the depth of every such intersection. They also mention a modification using topological sweep that achieves an $O(n^4)$ time and $O(n^2)$ space complexities, but some of the details remain unclear. The algorithms are based on the fact that the depth of a cell can be computed in constant time from the depth of an adjacent cell if the number of data points on both sides of the segment between the cells is known: with adjacent cells A, B, and a segment l between them, the simplicial depth of A equals the simplicial depth of B minus the number of simplicies containing B but not A (the number of points to the B side of l) plus the number of simplicies containing A, but not B (the number of data points to the A side of l).

Aloupis *et al.* algorithm initially calculates the number of data points strictly to one side of each of $O(n^2)$ segments in the arrangement of cells in $O(n^3)$ time and by calculating the depth of every data point in $O(n^2 \log n)$ time. For every segment l the algorithm finds and sorts all intersection points. It moves along l from one intersection point to the next, starting at an endpoint, and computes in constant update time the simplicial depth of each intersection. For each of $O(n^2)$ segments, all intersection points are sorted in $O(n^2 \log n)$ time, yielding an $O(n^4 \log n)$ time algorithm. The space complexity is $O(n^2)$. Another approach computes the arrangement of segments in $O(n^4)$ time and space and performs the same type of calculations. The authors suggest using topological sweep (which performs a sweep on an arrangement with m lines in $O(m^2)$ time and linear space) to yield an $O(n^4)$ time and $O(n^2)$ space algorithm. Since topological sweep is designed for an arrangement of lines, not segments, details about the implementation of the algorithm are needed. To adjust the topological sweep algorithm to this case, one would either need to extend the segments into lines, retaining information separating a pertinent segment from its extension, or need to generalize the concept of horizon trees in order to sweep the arrangement of segments instead of lines, allowing insertion and deletion of segments, at amortized constant update time.

According to the revised definition, the simplicial depth achieves its maximum value inside a cell, instead of on its boundary (Corollary 1). The simplicial median can be computed for the revised definition by a slight adjustment to the algorithms, maintaining the same time and space complexities. In the modified version the algorithm will advance between adjacent cells (instead of along segments) computing the depth of every cell. The algorithm can be further improved using halfspace depth (see Section 4).

4 Median and Bounds on Depth

To improve the constant factor on the method designed by Aloupis *et al.* [2], we use the halfspace depth to bound the simplicial depth.

Definition 5. Halfspace (Location) Depth for the sample version [10, 17]: Given a data set $\mathcal{S} = \{X_1, \ldots, X_n\}$, the *halfspace depth* of a point x with respect to $\mathcal{S}$ is the smallest number of data points contained in a closed half-plane passing through x.

As with simplicial depth (Lemma 1), the halfspace depth of all positions in a single cell is equal.

To the best of our knowledge no prior work was done on the relation between simplicial depth and halfspace depth in this context. Aloupis *et al.* [2] use the halfspace depth to reduce the time complexity of certain steps in their algorithm described in Section 3.3 (calculating the number of points that are strictly to one side of each segment and computing the simplicial depth of points), but their method does not reduce the complexity of the most time consuming step: traversing the $O(n^4)$ size arrangement. We describe how to compute bounds on the possible values for the simplicial depth of the cell using its halfpsace depth h and how these bounds can be used to improve the factor in the traversal step in the algorithm due to Aloupis *et al.* (Section 3.3).

Assume a data set $\mathcal{S} = \{X_1, \ldots, X_n\}$ in general position and add a query position x_o such that no two data points are collinear with it. As the set of points which cannot be chosen consists of a set of Lebesgue measure zero, and the interior of each cell has measure greater than zero, a position satisfying the above conditions exists in every cell. As x_o has half-space depth h, h is a lower bound on the number of data points in any halfplane defined by a line through x_o, and there exists a line l_1, passing through x_o, which partitions the data points into two sets such that exactly h points of the data set lie in one of them. By the Ham-Sandwich Cut theorem there exists a line l_2, passing through x_o, which partitions the data points into two sets such that one contains exactly $\lfloor \frac{n}{2} \rfloor$ data points, providing an upper bound on the size of any data set defined by a line through x_o. Both l_1 and l_2 can be chosen not to pass through any points of the data set.

Label the data points by the positive angle, from 0 to π, formed by the positive x-axis and the line which passes through the data point and x_o and sort according to this label. Consider two adjacent data points in the sorted list x_i and x_{i+1} and two more data points, x_j and x_k such that $\{j, k\} \cap \{i, i+1\} = \emptyset$. Then as x_i and x_{i+1} are adjacent in the sorted list, if both x_j and x_k are on the same half-plane defined by the line passing through x_i and x_o, they are also both on the same half-plane defined by the line passing through x_{i+1} and x_o. This implies that if the line passing through x_i and x_o divides the plane into two open half-planes, with α_1 data points in one open half-plane and α_2 data points in the other half-plane, the line passing through x_{i+1} and x_o divides the plane into two open half-planes with either α_1 data points in one open half-plane and α_2 data points in the other or $\alpha_1 \pm 1$ data points in one open half-plane and $\alpha_2 \mp 1$ data points in the other.

The simplicial depth of x_o, a position inside a cell, is given by Equation 1. There must be at least two points which divide the data set into two open half-planes with h and $n - h - 1$ data points, as well as at least two points which divide the data set into two open half-planes with $\frac{n}{2} - 1$ and $\frac{n}{2}$ data points for an even n (or $\frac{n}{2} - 1$ for an odd n). Additionally, for every $h \leq \alpha \leq \frac{n}{2} - 1$, there are at least two data points which divide the data set into two open half-planes with α points in one open half-plane and $n - \alpha - 1$ data points in the other half-plane. These restrictions apply to $n - 2h$ dividing half-planes so there are $2h$ half-planes remaining to consider.

Consider first the case where n is even. The simplicial depth value is maximized when h_i for the remaining dividing planes is as small as possible and is minimized when h_i is as large as possible; h_i is largest when all of the remaining $2h$ dividing planes have h data points in one open half-plane and $n - h - 1$ data points in the other. Similarly, it is minimized when all of the remaining $2h$ dividing half-planes contain $\frac{n}{2} - 1$ and $\frac{n}{2}$ data points producing the following bounds:

$$Max_{SD} = \binom{n}{3}^{-1}\left(\binom{n}{3} - \left(h\binom{\frac{n}{2}-1}{2} + h\binom{\frac{n}{2}}{2} + \sum_{i=h}^{\frac{n}{2}-1}\left(\binom{i}{2} + \binom{n-1-i}{2}\right)\right)\right)$$
$$= \binom{n}{3}^{-1}\left(\frac{h^3}{3} - \frac{h}{3} - \frac{nh^2}{2} + \frac{n^2h}{4}\right)$$
$$Min_{SD} = \binom{n}{3}^{-1}\left(\binom{n}{3} - \left(h\binom{h}{2} + h\binom{n-1-h}{2} + \sum_{i=h}^{\frac{n}{2}-1}\left(\binom{i}{2} + \binom{n-1-i}{2}\right)\right)\right)$$
$$= \binom{n}{3}^{-1}\left(\frac{-2h^3}{3} - h^2 - \frac{h}{3} + \frac{nh^2}{2} + \frac{nh}{2}\right)$$

For odd n there is an odd number of dividing planes with $\frac{n-1}{2}$ data points in each:

$$Max_{SD} = \binom{n}{3}^{-1}\left(\binom{n}{3} - \left((2h+1)\binom{\frac{n-1}{2}}{2} + \sum_{i=h}^{\frac{n-3}{2}}\left(\binom{i}{2} + \binom{n-1-i}{2}\right)\right)\right)$$
$$= \binom{n}{3}^{-1}\left(\frac{h^3}{3} - \frac{h}{12} - \frac{nh^2}{2} + \frac{n^2h}{4}\right)$$
$$Min_{SD} = \binom{n}{3}^{-1}\left(\binom{n}{3} - \left(h\binom{h}{2} + h\binom{n-1-h}{2} + \binom{\frac{n-1}{2}}{2} + \sum_{i=h}^{\frac{n-3}{2}}\left(\binom{i}{2} + \binom{n-1-i}{2}\right)\right)\right)$$
$$= \binom{n}{3}^{-1}\left(\frac{-2h^3}{3} - h^2 - \frac{h}{3} + \frac{nh^2}{2} + \frac{nh}{2}\right)$$

These bounds are monotonically increasing in the region of interest, where $0 \leq d \leq \frac{n}{2}$.

A simple construction shows that these bounds are tight. Consider the origin and 3 lines passing through the origin, which divide the plane into 6 slices. To bound the minimum, keep 3 of the slices empty, place alternating points in one pair of opposite slices (alternating, when considering the angle that the line through the origin and the point makes), and place enough points in the last slice to bring the number of points in the half-planes up to $\frac{n}{2}$. To get the maximum, keep 2 slices empty, put alternating points in two opposite cells and then for the two slices adjacent to one of these slices, place points such that the dividing planes will contain increasing number of data points, up to the $\frac{n}{2}$ level, and the other a decreasing number.

Using these bounds, we can improve the algorithm suggested by Aloupis *et al.* [2] (Section 3.3) and reduce the number of cells which could contain the simplicial depth median. The improvement is based on the fact that only cells with certain halfspace depth values (over a certain value that is computed by the algorithm) can achieve the maximal simplicial depth value.

To improve the algorithm start, as before, by finding the depth of all the data points. The deepest simplicial depth so far, $CurMax_{SD}$ is set to the maximal simplicial depth. Next, the halfspace depth d of the deepest non-trivial half-space contour is found, in $O(n \log^5 n)$ time and the minimum simplicial depth of a position with halfspace depth d, $Min_{SD}(d)$, is computed via the formulae above. If $Min_{SD}(d)$ is greater than $CurMax_{SD}$ then $CurMax_{SD}$ is updated. Finally, to find the halfspace depth contours that can achieve the median simplicial depth, the equation $Max_{SD}(d') - CurMax_{SD} = 0$ is solved for d. Then all positions with halfspace depth greater than or equal to d' could possibly be the median. The only cells which need to be checked for the median, due to the monotonicity of the minimum and maximum functions, using Aloupis *et al.*'s algorithm, are cells with halfspace depth between the bounds $\lceil d' \rceil$ and the halfspace median. As the halfspace depth of $\frac{n}{3}$ is always attained, plugging in the value of $Min_{SD}\left(\frac{n}{3}\right)$ implies that $d' \approx \frac{n}{2} - \frac{2}{n\sqrt[3]{7}} - \frac{n\sqrt[3]{7}}{6}$ and only $\frac{3n}{20}$ halfspace levels need to be checked.

To improve the $O(n^4)$ time and space algorithm, the arrangement is constructed, as before, but only a subset of the cells need to be traversed (those inside the contour with depth $\geq \lceil d' \rceil$). To improve the algorithm that achieves $O(n^4 \log n)$ time and $O(n^2)$ space complexities, note that only the segments that stab the contour of depth $\lceil d' \rceil$ need to be checked. Therefore, in the second step of the algorithm, only the subset of segments that stabs this contour is traversed.

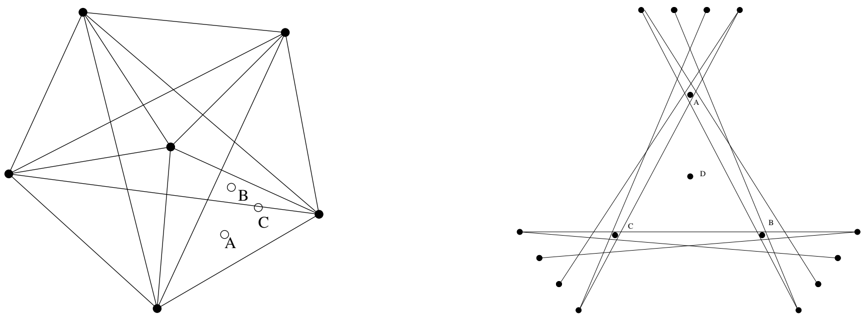
5 Approximate Algorithm

As there are currently no efficient algorithms to find simplicial depth contours in high dimensions, we propose a method for approximating the contours. As shown in Proposition 2, we can find local information about the simplicial depth function efficiently: specifically, we can find a discretized version of *the gradient*, the vector where simplicial depth of positions is increasing most rapidly. Using $O(n)$ observations, and thus finding $O(n)$ gradients, we can find a function to approximate the simplicial depth function, such as a multi-dimensional Taylor polynomial which minimizes the error in the gradient. The contours of this polynomial would be easier to find and would approximate the contours of the simplicial depth function.

6 Open Problems

6.1 Maximality and Monotonicity

Although the revised definition for simplicial depth fixes many of the examples presented by Zuo and Serfling [18], it does not achieve all desired properties in the sample case. Figure 3(b) shows an example where the data set has a unique center, D, but it neither attains maximality at the center, nor does it have monotonicity relative to the deepest point (properties P2 and P3).



(a) **(b)**

Figure 3: (a) Discontinuities in Simplicial depth at positions on facets: Choose 6 data points in $\mathbb{R}^2$ such that 5 are positioned evenly on a circle and 1 lies in the center of the circle. Any position A in one of the 5 outer cells has depth $\frac{4}{20}$ while any position B in a neighbor cell has depth $\frac{6}{20}$. Any position C on the boundary between the two cells is covered by 3 simplicies and on the boundary of 4 more and has depth $\frac{7}{20}$, according to Liu's definition, greater then A and B's. (b) A problem with the revised definition. The data points A,B, and C all have depth $(241 + \frac{1}{2}\binom{15}{2})/\binom{16}{3} = \frac{587}{1120}$ and the data point D, which is at the unique center of the data set has depth $(5^3 + \frac{1}{2}\binom{15}{2})/\binom{16}{3} = \frac{355}{1120}$. For clarity reasons not all cells are drawn.

6.2 Data Points

The revised definition does not solve all the problems in sample data sets, for instance, data points in $\mathbb{R}^2$ are over-counted as they are inherently a vertex of $\binom{n-1}{2}$ simplicies, whereas edges are inherently counted only $n-2$ times. This implies that the weight of a data point should be some factor λ of $\frac{1}{n-1}$, as $\frac{1}{2}(n-2) = \lambda\binom{n-1}{2} \Rightarrow \lambda = \frac{1}{n-1}$.

However, this factor isn't enough: consider a data set of n points, where $n-1$ points are evenly distributed angularly around one point. Then the depth of the center point should be at least as large as the depth of the n cells which use it as a vertex; however, this is not guaranteed by the $\frac{1}{n-1}$ factor. Thus the depth of a data point in $\mathbb{R}^d$ should be based both on the $\frac{1}{n-1}$ factor and the geometry of the $n-1$ other data points. Additionally, assuming invariance under dimension change (property P5), consider a simplex in $\mathbb{R}^d$, then it is a facet of $n-d-1$ $(d+1)$-dimensional simplicies. To maintain the ratio ρ between the two dimensions, we get $\binom{n}{d+1}^{-1} = \binom{n}{d+2}^{-1}\frac{1}{2}(n-d-1)\,\rho \Rightarrow \rho = \frac{2}{d+2}$. For a given k-dimensional simplex, $0 \le k \le d$, which is part of $\binom{n-k-1}{d-k}$ d-dimensional simplicies with weight w_1 and part of $\binom{n-k-1}{d+1-k}$ $(d+1)$-dimensional simplicies with weight w_2, we have $\binom{n}{d+1}^{-1}\binom{n-k-1}{d-k}w_1 = \binom{n}{d+2}^{-1}\left(\frac{2}{d+2}\right)\binom{n-k-1}{d+1-k}w_2 \Rightarrow w_1 = \frac{1}{2}(d+1-k)w_2$. Thus, to guarantee property P5, the weight of a point in $\mathbb{R}^2$, and more generally, any h-dimensional simplex, should not only have a constant multiple for the weight, but also take into consideration the geometry of the data set, similarly to the properties for the revised definition described in Section 2.1.

7 Summary

We present a modification to the definition of simpilicial depth, that solves some of the problems raised in the past. We are currently investigating how to cope with the computation of the depth of data points in high dimensions, while maintaining the desirable properties, as described in Section 6.2. In addition we are working on approximation algorithms based on the local properties of the depth function (Section 5), to enable efficient approximation for high dimensional data.

In addition, we presented a connection between simplicial depth and halfspace depth. We believe that this relation can be further utilized to study the properties of the two depth functions and further improve algorithms' complexity.

References

[1] G. Aloupis, C. Cortes, F. Gomez, M. Soss, and G. Toussaint. Lower bounds for computing statistical depth. *Computational Statistics & Data Analysis*, 40(2):223–229, 2002.

[2] G. Aloupis, S. Langerman, M. Soss, and G. Toussaint. Algorithms for bivariate medians and a Fermat-Torricelli problem for lines. In *Proc. 13th Canadian Conf. on Comp. Geo.*, pages 21–24, 2001.

[3] Imre Bárány. A generalization of Carathéodory's theorem. *Discrete Math.*, 40(2-3):141–152, 1982.

[4] V. Barnett. The ordering of multivariate data. *J. Roy. Statist. Soc. Ser. A*, 139(3):318–355, 1976.

[5] E. Boros and Z. Füredi. The number of triangles covering the center of an n-set. *Geom. Dedicata*, 17(1):69–77, 1984.

[6] A. Y. Cheng and M. Ouyang. On algorithms for simplicial depth. In *Proc. 13th Canadian Conference on Computational Geometry*, pages 53–56, 2001.

[7] W. Eddy. Convex hull peeling. In H. Caussinus, editor, *COMPSTAT*, pages 42–47. Physica-Verlag, Wien, 1982.

[8] H. Edelsbrunner. *Algorithms in combinatorial geometry*, volume 10 of *EATCS Monographs on Theoretical Computer Science*. Springer-Verlag, Berlin, 1987.

[9] J. Gil, W. Steiger, and A. Wigderson. Geometric medians. *Discrete Math.*, 108(1-3):37–51, 1992. Topological, algebraical and combinatorial structures. Frolík's memorial volume.

[10] J. Hodges. A bivariate sign test. *The Annals of Mathematical Statistics*, 26:523–527, 1955.

[11] S. Khuller and J. S. B. Mitchell. On a triangle counting problem. *Inform. Process. Lett.*, 33(6):319–321, 1990.

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

[13] R. Liu, J. Parelius, and K. Singh. Multivariate analysis by data depth: descriptive statistics, graphics and inference (with discussions). *The Annals of Statistics*, 27:783–858, 1999.

[14] P. C. Mahalanobis. On the generalized distance in statistics. In *Proceedings of the National Academy India*, volume 12, pages 49–55, 1936.

[15] P.J. Rousseeuw and I. Ruts. Bivariate location depth. *Applied Statistics-Journal of the Royal Statistical Society Series C*, 45(4):516–526, 1996.

[16] K. Singh. On the majority data depth. Technical report, Dep. of Statistics, Rutgers University, 1993.

[17] J.W. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), Vol. 2*, pages 523–531. Canad. Math. Congress, Montreal, Que., 1975.

[18] Y. Zuo and R. Serfling. General notions of statistical depth function. *Ann. Statist.*, 28(2):461–482, 2000.