Approximating the Simplicial Depth in High Dimensions

Peyman Afshani* Donald R. Sheehy† Yannik Stein‡

Abstract

Let P be a set of n points in d-dimensions. The simplicial depth $\sigma_P(q)$ of a point q is the number of d-simplices with vertices in P that contain q in their convex hulls. The simplicial depth is a notion of data depth with many applications in robust statistics and computational geometry. Computing the simplicial depth of a point is known to be a challenging problem. The trivial solution requires $O(n^{d+1})$ time whereas it is generally believed that one cannot do better than $O(n^{d-1})$.

We present two approximation algorithms for computing the simplicial depth of a point in high dimensions with different worst-case scenarios. By combining these approaches, we can compute a $(1+\varepsilon)$-approximation of the simplicial depth in time $\tilde{O}(n^{d/2+1})$ with high probability ignoring polylogarithmic factors. Furthermore, we present a simple strategy to compute the simplicial depth exactly in $O(n^d \log n)$ time, which provides the first improvement over the trivial $O(n^{d+1})$ time algorithm for $d > 4$. Finally, we show that computing the simplicial depth exactly is #P-complete and W[1]-hard if the dimension is part of the input.

1 Introduction

Let $P \subset \mathbb{R}^d$ be a point set and $q \in \mathbb{R}^d$ be a point. The *simplicial depth* [14] $\sigma_P(q)$ of q with respect to P is the number of subsets $P' \subseteq P$, $|P'| = d+1$, that contain q in their convex hull (see also [4] for an alternate definition). This is one of the important definitions of data depth and has generated interest in both robust statistics and computational geometry since its introduction. Designing efficient algorithms to compute (or approximate) the simplicial depth of a point remains an intriguing task in this area.

Computing the simplicial depth of a single point in 2D was considered even before its formal definition [11] almost three decades ago, perhaps because it translates into an "intuitive" problem of counting the number of triangles containing a given point. In fact, at least three independent papers study this problem in 2D and show how to compute the simplicial depth in $O(n \log n)$ time [9, 11, 14]. This running time is optimal [1]. In 2003, Burr et al. [4] presented an alternate definition for the simplicial depth to overcome some unpleasant behaviors that emerge when dealing with degeneracies. Since we will be dealing with approximations, we will assume general position and thus avoid issues with degeneracy. In 3D, the first non-trivial result offered the bound of $O(n^2)$ [9] but it was flawed; fortunately, the running time of $O(n^2)$ could still be obtained with proper modifications [7]. The same authors presented an algorithm with running time of $O(n^4)$ in 4D. For dimensions beyond 4 there seems to be no significant improvements over the trivial $O(n^{d+1})$ brute-force solution. Furthermore, it is natural to conjecture that computing the simplicial depth should require $\Omega(n^{d-1})$ time: given a set P of n points, it is generally conjectured that detecting whether or not $d+1$ points lie on a hyperplane requires $\Omega(n^d)$ time [8] and this conjecture would imply that detecting whether d points of P and a fixed point q lie on a hyperplane should require $\Omega(n^{d-1})$ time. This is one motivation to consider the approximate version of the problem. In fact, Burr et al. [4] have already expressed interest in computing an approximation to the simplicial depth and they propose a potential approach, although without any worst-case analysis [3].

Here, we only consider relative approximation; additive approximation (with additive error of εn^{d+1}) can be obtained using ε-nets and ε-approximations (see [5, 2] for more details).

Another motivation for computing a relative approximation comes from applications in outlier removal. Intuitively, statistical depth measures how deep a point is embedded in the data cloud with outliers corresponding to points with small values of depth. In such applications, if a small relative error of $(1+\varepsilon)$ is tolerable, then faster outlier removal can be possible using approximations.

2 Approximation in High Dimensions

In this section, we present two approximation algorithms for simplicial depth in high dimensions, each with a different worst case scenario. By combining these strategies, we obtain a constant factor approxi-

*MADALGO, Department of Computer Science, Aarhus University, Denmark, peyman@madalgo.au.dk. Supported in part by the Danish National Research Foundation grant DNRF84 through Center for Massive Data Algorithmics (MADALGO).

†University of Connecticut, USA, don.r.sheehy@gmail.com.

‡Institut für Informatik, Freie Universität Berlin, Germany, yannik.stein@fu-berlin.de. Supported by the Deutsche Forschungsgemeinschaft within the research training group "Methods for Discrete Structures" (GRK 1408).

mation algorithm with $\tilde{O}(n^{d/2+1})$ running time.

2.1 Small Simplicial Depth: Enumeration

Let $P \subset \mathbb{R}^d$ be a set and $q \in \mathbb{R}^d$ a query point. We denote with Δ_P the set of all d-simplices with vertices in P. If $\sigma_P(q)$ is small, a simple counting approach that iterates through all simplices $\Delta \in \Delta_P$ leads to an efficient algorithm. The key is to construct a graph that contains exactly one node per simplex $\Delta \in \Delta_P$. Then, counting can be carried out by a breadth-first search and we avoid looking at subsets of P that do not contain q in their convex hull. For this, we use the Gale transform to dualize the problem. We shortly restate important properties of the Gale transform. For more details see [13]. Let in the following $\mathbf{0}$ denote the origin.

Lemma 1 *Let $P = \{p_1, \ldots, p_n\} \subset \mathbb{R}^d$ be a point set with $\sigma_P(\mathbf{0}) > 0$. Then, there is a set $\bar{P} = \{\bar{p}_1, \ldots, \bar{p}_n\} \subset \mathbb{R}^{n-d-1}$ such that a $(d+1)$-subset $P' \subseteq P$ contains $\mathbf{0}$ in its convex hull iff $\bar{P} \setminus \{\bar{p}_i \mid p_i \in P'\}$ defines a facet of $\operatorname{conv}(\bar{P})$.*

Consider now the graph $G_P(q) = (V, E)$ with $V = \Delta_P$. Two simplices Δ, Δ' are adjacent iff Δ' can be obtained from Δ by swapping one point in Δ with a different point in P. We call $G_P(q)$ the *simplicial graph* of P with respect to q.

Lemma 2 *Let $P \subset \mathbb{R}^d$ be a set of size n. Then, $G_P(q)$ is $(n-d-1)$-connected and $(n-d-1)$-regular.*

Proof. We assume w.l.o.g. that $q = \mathbf{0}$. Let Δ, Δ' be two adjacent nodes in $G_P(q)$. Furthermore let $\bar{P}$ denote the Gale transform of P. Set $\bar{\Delta} = \{\bar{p} \mid p \in P \setminus \Delta\}$ and $\bar{\Delta}' = \{\bar{p} \mid p \in P \setminus \Delta'\}$. By Lemma 1, the two sets $\bar{\Delta}$ and $\bar{\Delta}'$ define facets of $\operatorname{conv}(\bar{P})$. Since Δ and Δ' are adjacent, we have $|\Delta \cap \Delta'| = d$ and hence $|\bar{\Delta} \cap \bar{\Delta}'| = n - d - 2$. Thus, the facets defined by $\bar{\Delta}$ and $\bar{\Delta}'$ share a ridge. Hence, $G_P(q)$ is isomorph to the 1-skeleton of the polytope dual to $\operatorname{conv}(\bar{P})$. In particular, this implies that $G_P(q)$ is $(n-d-1)$-connected. It remains to show that the graph is $(n-d-1)$-regular. Let $\Delta \in V$ be a node. It is easy to see that each of the $n-d-1$ points in $P \setminus \Delta$ can be swapped in, each time resulting in a distinct simplex. ☐

Since $G_P(q)$ is connected, we can count the number of vertices using BFS.

Lemma 3 *Let $P \subset \mathbb{R}^d$ be a set of size n and $q \in \mathbb{R}^d$ a query point. Then, $\sigma_P(q)$ can be computed in $O(n\sigma_P(q))$ time.*

2.2 Large Simplicial Depth: Sampling

If the simplicial depth is large, the enumeration approach becomes infeasible. In this case we apply a simple random sampling algorithm.

Lemma 4 *Let $P \subset \mathbb{R}^d$ be a set and $q \in \mathbb{R}^d$ a query point. Furthermore, let $\varepsilon, \delta > 0$ be constants and let $m \in \mathbb{N}$ be a parameter. If $\sigma_P(q) \geq m$, then $\sigma_P(q)$ can be $(1+\varepsilon)$-approximated in $\tilde{O}(n^{d+1}/m)$ time with error probability $O(n^{-\delta})$.*

Proof. Let $\Delta_1, \ldots, \Delta_k$ be k random $(d+1)$-subsets of P for $k = \left\lceil \frac{4\delta n^{d+1} \log n}{\varepsilon^2 m} \right\rceil$. For each random subset Δ_i, let X_i be 1 iff $q \in \operatorname{conv}(\Delta_i)$ and 0 otherwise. We have $\mu = \mathbb{E}[\sum_{i=1}^k X_i] = k\frac{\sigma_P(q)}{n^{d+1}} = \frac{4\delta\sigma_P(q)\log n}{\varepsilon^2 m} \geq \frac{4\delta}{\varepsilon^2}\log n$. Applying the Chernoff bound, we get $\Pr[|\sum_{i=1}^k X_i - \mu| \geq \varepsilon\mu] = O(n^{-\delta})$. Thus, $\frac{n^{d+1}}{k}X$ is a $(1+\varepsilon)$-approximation of $\sigma_P(q)$ with error probability $O(n^{-\delta})$.

For $d = O(1)$, we can test in $O(1)$ whether a given $(d+1)$-subset of P contains a point in its convex hull. Hence, the running time is dominated by the number of samples. ☐

2.3 Combining the Strategies

Theorem 5 *Let $P \subset \mathbb{R}^d$ be a set and $q \in \mathbb{R}^d$ a query point. Furthermore, let $\varepsilon > 0$ and $\delta > 0$ be constants. Then, $\sigma_P(q)$ can be $(1+\varepsilon)$-approximated in $\tilde{O}(n^{d/2+1})$ time with error probability $O(n^{-\delta})$.*

Proof. We apply the algorithm from Lemma 3 and stop it once $n^{d/2}$ nodes of $G_P(q)$ are explored. This requires $O(n^{d/2+1})$ time. If the graph is not yet fully explored, we know $\sigma_P(q) \geq n^{d/2}$. We can now apply the algorithm from Lemma 4 and compute a $(1+\varepsilon)$-approximation in $\tilde{O}(n^{d/2+1})$ time with error probability $O(n^{-\delta})$. ☐

3 An Exact Algorithm in High Dimensions

In this section we describe a simple strategy to compute the simplicial depth exactly in $O(n^d \log n)$ time. While we do not achieve the conjectured lower bound of $\Omega(n^{d-1})$, we cut down roughly a factor n compared to the trivial upper bound of $O(n^{d+1})$. Note that this almost matches the best previous bound of $O(n^4)$ in 4D as well [7].

W.l.o.g, assume q is the origin, $\mathbf{0}$. Our main idea is very simple: consider d points $p_1, \ldots, p_d \in P$. Let $\overrightarrow{r_i}$ be the ray that originates from $\mathbf{0}$ towards $-p_i$. We would like to count how many points $p \in P$ can create a simplex with $p_1, \ldots, p_d$ that contains $\mathbf{0}$. We observe that this is equivalent to counting the number of points of P that lie inside the simplex created by rays

$\overrightarrow{r_1}, \ldots, \overrightarrow{r_d}$. We can count this number in polylogarithmic time if we spend $\tilde{O}(n^d)$ time to build a simplex range counting data structure on P. This would give an algorithm with overall running time of $\tilde{O}(n^d)$. We can cut the log factors down to one by employing a slightly more intelligent approach.

We use the following observation made by Gil et al. [9].

Observation 1 *Let q be a point inside a simplex $a_1 \ldots a_{d+1}$ and let a'_i be a point on the ray $\overrightarrow{qa_i}$. Then, $q \in \mathrm{conv}\{a_1, \ldots, a_{i-1}, a'_i, a_{i+1}, \ldots, a_{d+1}\}$.*

Pick two arbitrary parallel hyperplanes h_1 and h_2 such that P lies between them. This can be done easily in $O(n)$ time. Next, using central projection from $\mathbf{0}$, we map the points onto the hyperplanes h_1 and h_2: for every point $p_i \in P$, we create the ray $\overrightarrow{\mathbf{0}p_i}$ and let p'_i be the intersection of the ray with h_1 or h_2. Thus, the point set P can be mapped to two point sets P_1 and P_2 where P_1 lies on h_1 and p_2 lies on P_2 and furthermore, by Observation 1, $\sigma_P(q) = \sigma_{P_1 \cup P_2}(q)$.

Now we use the following result from the simplex range counting literature.

Theorem 6 *[6] Given a set of n points in d-dimensional space, and any constant $\varepsilon > 0$, one can build a data structure of size $O(n^{d+\varepsilon})$ in $O(n^{d+\varepsilon})$ expected preprocessing time, such that given any query simplex Δ, the number of points in Δ can be counted in $O(\log n)$ time.*

We build the above data structure on P_1 and P_2. However, since both of these point sets lie on a $(d-1)$-dimensional flat, the preprocessing time is $O(n^{d-1+\varepsilon}) = O(n^d)$ if we choose $\varepsilon = 1/2$. Next, for any d tuples of points $p_1, \ldots, p_d$, we create the rays $\overrightarrow{r_1}, \ldots, \overrightarrow{r_d}$ and the corresponding simplex Δ. We find the intersection of Δ in $O(1)$ time with hyperplanes h_1 and h_2 and issue two simplex range counting queries, one in each hyperplane. Thus, in $O(\log n)$ time, we can count how many simplices contain $\mathbf{0}$ that are made by points $p_1, \ldots, p_d$. We add all these numbers over all d tuples, which counts each simplex containing $\mathbf{0}$ exactly $(d+1)$ times. The number of d-tuples is $O(n^d)$ and for each we spend $O(\log n)$ time querying the data structures. Thus, we obtain the following theorem.

Theorem 7 *Given a set P of n points in $\mathbb{R}^d$, the simplicial depth of a point p can be computed in $O(n^d \log n)$ expected time.*

4 Complexity

Let $P \subset \mathbb{R}^d$ be a set of n points and $q \in \mathbb{R}^d$ a query point. If the dimension is constant, then clearly computing $\sigma_P(q)$ can be carried out in polynomial time. We now consider the case that d is part of the input. We show that in this case computing the simplicial depth is #P-complete by a reduction from counting the number of perfect matchings in bipartite graphs.

Theorem 8 *Let $P \subset \mathbb{R}^d$ be a set and $q \in \mathbb{R}^d$ a query point. Then, computing $\sigma_P(q)$ is #P-complete if the dimension is part of the input.*

Proof. Let $G = (V, E)$ be a bipartite graph with $|V| = n$ and $|E| = m$. It is well known that computing the number of perfect matchings in G is #P-complete [15]. Let $\mathcal{P}_H \subset \mathbb{R}^m$ be the perfect matching polytope for G [10, Chapter 30]. It is defined by $m + 2n$ half-spaces. Furthermore, the number of vertices of $\mathcal{P}_H$ equals the number k of perfect matchings in G. Consider now the dual polytope $\mathcal{P}_V \subset \mathbb{R}^m$. It is the convex hull of $m + 2n$ points $P \subset \mathbb{R}^m$ and the number of facets equals k. Let $\bar{P} \subset \mathbb{R}^{2n-1}$ be the Gale transform of P. By Lemma 1, there is a bijection between the facets of $\mathcal{P}_V$ and the $(2n-1)$-simplices with vertices in $\bar{P}$ that contain $\mathbf{0}$ in their convex hull. Hence, $\sigma_{\bar{P}}(\mathbf{0}) = k$. □

Next, we show that computing the simplicial depth is W[1]-hard with respect to the parameter d by a reduction to *d-Carathéodory.* In d-Carathéodory, we are given a set $P \subset \mathbb{R}^d$ and have to decide whether there is a $(d-1)$-simplex with vertices in P that contains $\mathbf{0}$ in its convex hull. Knauer et al. [12] proved that this problem is W[1]-hard with respect to the parameter d.

Theorem 9 *Let $P \subset \mathbb{R}^d$ be a set and $q \in \mathbb{R}^d$ a query point. Then, computing $\sigma_P(q)$ is W[1]-hard with respect to the parameter d.*

Proof. Assume we have access to an oracle that, given a query point q and a set $Q \subset \mathbb{R}^d$, returns $\sigma_Q(q)$. We show that #d-Carathéodory can be decided with two oracle queries.

Let k_d denote the number of $(d-1)$-simplices with vertices in P that contain $\mathbf{0}$ in their convex hulls and let k_{d+1} denote the number of d-simplices with vertices in P that contain $\mathbf{0}$ in their interior. Then $\sigma_P(\mathbf{0})$ can be written as $(|P| - d)k_d + k_{d+1}$. We want to decide whether $k_d > 0$. For each point $p \in P$ let $\tilde{p} \in \mathbb{R}^{d+1}$ denote the $(d+1)$-dimensional point that is obtained by appending a 1-coordinate and similarly, for each subset $P' \subset P$ let $\tilde{P'}$ denote the set $\{\tilde{p} \mid p \in P'\} \subset \mathbb{R}^{d+1}$. We denote with S the set $\{(0, \ldots, 0, -1)^T, (0, \ldots, 0, -2)^T\} \subset \mathbb{R}^{d+1}$ and set $Q = \tilde{P} \cup S$. Again, we want to express $\sigma_Q(\mathbf{0})$ as a function of k_d and k_{d+1}. Let $Q' \subset Q$, $|Q'| = d+2$, be a subset that contains $\mathbf{0}$ in its convex hull. Clearly, Q' has to contain a point from S. Let $\tilde{P}' = Q' \cap \tilde{P}$ denote the part from $\tilde{P}$ and let $S' = Q' \cap S$ denote the part from S. By construction of S, we have

$(0,\ldots,0,1)^T \in \mathrm{conv}(\tilde{P}')$ and hence $\mathbf{0} \in \mathrm{conv}(P')$. That is, each $(d+2)$-simplex with vertices in Q that contains $\mathbf{0}$ in its convex hull corresponds to either a d-simplex or a $(d-1)$-simplex with vertices in P that contains $\mathbf{0}$ in its convex hull. Consider now a set $P' \subset P$ with $|P'| = d+1$ and $\mathbf{0} \in \mathrm{conv}(P)$. Then, the corresponding set $\tilde{P}'$ can be extended in two ways to a subset $Q' \subset Q$, $|Q'| = d+2$, with $\mathbf{0} \in \mathrm{conv}(Q')$ by taking either point in S. On the other hand, if $P' \subset P$ is a subset of size d with $\mathbf{0} \in \mathrm{conv}(P')$, then we can extend $\tilde{P}'$ to a set $Q' \subset Q$, $|Q'| = d+2$, with $\mathbf{0} \in \mathrm{conv}(Q')$ by either taking both points in S or by taking one arbitrary point in $\tilde{P} \setminus \tilde{P}'$ and either point in S. Hence, we have $\sigma_Q(\mathbf{0}) = 2k_{d+1} + k_{d-1} + 2(|P|-d)k_{d-1}$. Since $k_d = \sigma_Q(\mathbf{0}) - 2\sigma_P(\mathbf{0})$, we can decide whether $k_d > 0$ with two oracle queries. ☐

The following theorem is now immediate.

Theorem 10 *Let $P \subset \mathbb{R}^d$ be a set of d-dimensional points and $q \in \mathbb{R}^d$ a query point. Then, computing $\sigma_P(q)$ is #P-complete and W[1]-hard with respect to the parameter d.*

We conclude the section with a constructive result: although computing the simplicial depth is #P-complete, it is possible to determine the parity in polynomial-time.

Theorem 11 *Let $P \subset \mathbb{R}^d$ be a set of points and $q \in \mathbb{R}^d$ a query point. If $n-d-1$ is odd or $\binom{n}{d}$ is even, then $\sigma_P(q)$ is even. Otherwise, $\sigma_P(q)$ is odd.*

Proof. We assume w.l.o.g. that q is the origin. Since the simplicial graph $G_P(\mathbf{0})$ is $(n-d-1)$-regular, the product $(n-d-1)|V| = (n-d-1)\sigma_P(\mathbf{0})$ is even. If $(n-d-1)$ is odd, $\sigma_P(q)$ has to be even. Assume now $(n-d-1)$ is even. We construct a new point set Q in $\mathbb{R}^{d+1}$ similar as in the proof of Theorem 9. Let R denote the set $\{(0,\ldots,0,-1)^T, (0,\ldots,0,2)^T\} \subset \mathbb{R}^{d+1}$ and set $Q = \tilde{P} \cup R \subset \mathbb{R}^{d+1}$, where $\tilde{P}$ is defined as in the proof of Theorem 9. Let us now consider the graph $G_Q(\mathbf{0})$. Since $n-d-1$ is even, $(|Q|-(d+1)-1) = n-d$ is odd. Now, $G_Q(\mathbf{0})$ is $(n-d)$-regular and thus $\sigma_Q(\mathbf{0})$ is even. Let $Q' \subset Q$, $|Q| = d+2$, be a subset that contains the origin in its convex hull. Then either (i) $R \subset Q'$ or (ii) Q' contains the point $r = (0,\ldots,0,-1)^T \in R$ and $d+1$ points $\tilde{P}' \subseteq \tilde{P}$ with $(0,\ldots,0,1)^T \in \mathrm{conv}(\tilde{P}')$. There are $\binom{n}{d}$ sets Q' with Property (i) and $\sigma_P(\mathbf{0})$ sets Q' with Property (ii). Hence, we have $\sigma_Q(\mathbf{0}) = \sigma_P(\mathbf{0}) + \binom{n}{d}$ is even and thus $\sigma_P(\mathbf{0})$ is odd iff $\binom{n}{d}$ is odd. ☐

Acknowledgements. This work was initiated while YS was visiting MADALGO in Aarhus. I would like to thank the working group for their hospitality and for a constructive atmosphere.

References

[1] G. Aloupis, C. Cortés, F. Gómez, M. Soss, and G. Toussaint. Lower bounds for computing statistical depth. *Comput. Stat. & Data Anal.*, 40(2):223–229, 2002.

[2] Amitabha Bagchi, Amitabh Chaudhary, David Eppstein, and Michael T Goodrich. Deterministic sampling and range counting in geometric data streams. *TALG 2007*, 3(2):16.

[3] Michael Burr, Eynat Rafalin, and Diane L. Souvaine. Simplicial depth: An improved definition, analysis, and efficiency for the finite sample case.

[4] Michael Burr, Eynat Rafalin, and Diane L. Souvaine. Simplicial depth: An improved definition, analysis, and efficiency for the finite sample case. In *CCCG*, pages 136–139, 2004.

[5] B. Chazelle. *The discrepancy method: randomness and complexity.* Cambridge University Press, 2000.

[6] Bernard Chazelle, Micha Sharir, and Emo Welzl. Quasi-optimal upper bounds for simplex range searching and new zone theorems. *Algorithmica*, 8:407–429, 1992.

[7] A. Y. Cheng and M. Ouyang. On algorithms for simplicial depth. In *CCCG*, pages 53–56, 2001.

[8] Jeff Erickson. New lower bounds for convex hull problems in odd dimensions. *SIAM Journal on Computing*, 28(4):1198–1214, 1999.

[9] J. Gil, W. Steiger, and A. Wigderson. Geometric medians. *Discrete Math.*, 108(1):37–51, 1992.

[10] R. L. Graham, M. Grötschel, and L. Lovász. *Handbook of combinatorics*, volume 2. Elsevier, 1995.

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

[12] C. Knauer, H. R. Tiwary, and D. Werner. On the computational complexity of Ham-Sandwich cuts, Helly sets, and related problems. In *STACS 2011*, pages 649–660.

[13] R. Thomas R. *Lectures in Geometric Combinatorics.* American Mathematical Society, 2006.

[14] P. J. Rousseeuw and I. Ruts. Bivariate location depth. *J. R. Stat. Soc. Series C (Appl. Stat.)*, 45(4):516–526, 1996.

[15] L. G. Valiant. The complexity of computing the permanent. *Theor. Comput. Sci.*, 8(2):189 – 201, 1979.