

# Computing Batched Depth Queries and the Depth of a Set of Points

by

Sachini Kanchana Rajapakse

A Thesis submitted to the Faculty of Graduate Studies of  
The University of Manitoba  
in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Department of Computer Science  
University of Manitoba  
Winnipeg

Copyright © 2022 by Sachini Kanchana Rajapakse

# Computing Batched Depth Queries and the Depth of a Set of Points

Thesis advisors

Dr. Stephane Durocher

Dr. Alexandre Leblanc

Author

Sachini K. Rajapakse

## Abstract

In Computational Geometry and Statistics, we often encounter problems related to data depth. Data depth describes the centrality of an object relative to a set of objects. Simplicial depth and Tukey depth are two such common depth measures for expressing the depth of a point  $q$  relative to a set  $P$  of points in  $\mathbb{R}^d$ . In this thesis, we discuss the problems of how efficiently we can compute the depths of  $k$  query points in a batch relative to a set  $P$  of  $n$  points. We examine three algorithms for computing batched simplicial depth queries and two algorithms to compute batched Tukey depth queries in  $\mathbb{R}^2$ . Depending on the relative cardinalities of  $P$  and  $Q$ , one algorithm performs better than others for each depth measure. Then, we introduce new notions to express the depth of a set  $Q$  of points relative to a set  $P$  of points in  $\mathbb{R}^d$  as the average depth of points in  $Q$ . We discuss how to compute these new notions of depth in  $\mathbb{R}^2$  by applying the algorithms above, giving an algorithm for computing the simplicial depth of a set  $Q$  relative to set  $P$  in  $O(\min\{kn \log n, n^2 + nk, n^4 + k \log n\})$  time, and the Tukey depth of a set  $Q$  relative to set  $P$  in  $O(\min\{kn \log n, n^2 + k \log n\})$  time, when  $n = |P|$  and  $k = |Q|$ . Further, we consider different statistical and probabilistic interpretations of these new notions of depth of a set. We show that the simplicial depth of  $Q$  relative to  $P$  is proportional to the expected number of points of  $Q$  contained in a simplex constructed from points selected randomly from  $P$ . Properties of depth measures are commonly analyzed to compare and contrast different depth measures, which are initially introduced for the depth measures of a single query point; we present generalizations for five of these properties for the depth measures of sets and evaluate these properties to compare depth measures defined for a set of points.

# Acknowledgement

First and foremost, I would like to thank my supervisor Dr. Stephane Durocher, for providing me the privilege to pursue graduate studies. He has been not only an incredible mentor but also an inspiring role model to learn from. I would also like to thank my co-supervisor, Dr. Alexandre Leblanc, for his continuous support and guidance throughout my studies. I am genuinely grateful to my supervisors for how they guided me to look at research problems, inspired me to explore various fields of research, and also for their generosity with time and resources as well as their patience, especially during a challenging time of the pandemic. It has been a great honor and privilege for me to work under their supervision.

I would also like to thank the members of my advisory committee, Dr. Karen Gunderson and Dr. Max Turgeon for accepting to be a member of my advisory committee. Their valuable advice and academic support assisted me to conduct this thesis.

I am truly grateful to the funding assistance by the Natural Sciences and Engineering Research Council of Canada (NSERC), the University of Manitoba Graduate Fellowship (UMGF) and scholarship programs, Faculty of Graduate Studies Research Completion Scholarship, Reine-Baniuk Memorial Scholarship that financially supported me to complete my graduate studies.

Last but not least, I would like to express my deep gratitude to my family and friends. Special thanks to my wonderful husband, Charith. I could not have gotten through this without their amazing support and understanding of them. Thanks all for always being there and believing in me.

*“For my beloved family  
who always believes in me and supports me.  
For all my teachers and all the taxpayers in Sri Lanka  
who supported my free education”.*

# Contents

<b>Abstract</b>	<b>i</b>
<b>Acknowledgement</b>	<b>i</b>
<b>Dedication</b>	<b>ii</b>
<b>Contents</b>	<b>v</b>
<b>List of Figures</b>	<b>viii</b>
<b>List of Tables</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Our Main Results . . . . .	3
1.2 Thesis Outline . . . . .	6
<b>2 Literature Review</b>	<b>7</b>
2.1 Data Depth . . . . .	7
2.1.1 Tukey Depth . . . . .	8
2.1.2 Simplicial Depth . . . . .	9
2.2 Depth-Related Medians . . . . .	10
2.2.1 Tukey Depth Median and Contours . . . . .	11
2.2.2 Simplicial Median . . . . .	11
2.3 Depths of Multiple Query Points . . . . .	13
2.3.1 Simplicial Depth of a Query Set . . . . .	14
2.3.2 Tukey Depth of a Query Set . . . . .	15
2.3.3 Depth Histograms . . . . .	16

<b>3</b>	<b>Computing the Depths of Batched Query Points</b>	<b>19</b>
3.1	Algorithms for Computing Simplicial Depths of Batched Query Points	20
3.1.1	Algorithm S.I . . . . .	20
3.1.2	Algorithm S.II . . . . .	20
3.1.3	Algorithm S.III . . . . .	23
3.2	Algorithms for Computing Tukey Depths of Batched Query Points . .	25
3.2.1	Algorithm T.I . . . . .	26
3.2.2	Algorithm T.II . . . . .	26
3.3	Application of Batched Depth Queries . . . . .	27
<b>4</b>	<b>Defining the Depth of a Query Set</b>	<b>29</b>
4.1	Defining the Simplicial Depth of a Query Set . . . . .	29
4.2	Probabilistic Interpretation for the Simplicial Depth of a Query Set .	34
4.3	Tukey Depth of a Query Set . . . . .	37
4.4	Applications of Depth of Sets . . . . .	40
<b>5</b>	<b>Properties for Depth Measures of Sets of Points</b>	<b>43</b>
5.1	P1 - Affine Invariance . . . . .	44
5.2	P2 - Median at the Center of Symmetry . . . . .	45
5.3	P3 - Monotonicity Relative to the Deepest Point . . . . .	46
5.4	P4 - Vanishing at Infinity . . . . .	47
5.5	P5 - Consistency Across Dimensions . . . . .	49
5.6	P6 - Convexity of Depth Contours . . . . .	49
5.7	P7 - Breakdown Point . . . . .	50
5.8	Summary of Properties . . . . .	50
<b>6</b>	<b>Conclusion and Directions for Future Research</b>	<b>52</b>
6.1	Batched Depth Queries . . . . .	52
6.1.1	Dynamic Simplicial Depth . . . . .	53
6.2	Depth of a Query Set of Points . . . . .	55
6.3	Properties Defined for Depth Measures of Sets of Points . . . . .	57
	<b>Bibliography</b>	<b>63</b>

# List of Figures

2.1	Simplicial depth calculation . . . . .	17
3.1	Points in the primary plane $(q, p_1 \dots p_8)$ are transformed to the set of lines $(L_q, L_1 \dots)$ in the dual plane. Here $O$ is the origin and the line connecting origin and $q$ ( $Oq$ line) partitions the primal plane into four quadrants at $q$ . While finding the intersection order of $L_P$ lines in $L_q$ , insert $p_i$ to the list of the relevant quadrant. Combining four lists gives the angular order of $P$ around $q$ . This figure replicates an example illustrated by Lee and Ching [27, Figure 3]. . . . .	22
3.2	The arrangement $L$ as described with $n = 5$ points. Traversing $L$ for computing the depth of each cell can be started from outside the convex hull of $P$ with zero depth. The depth of a neighboring cell is computed by subtracting the number of points to the left of the shared edge and adding the number of points to the right of the shared edge. For example, to find simplicial depth of $C_7$ use the simplicial depth of $C_6$ which share common blue edge $e$ and has know simplicial depth $d = 4$ . Now, simplicial depth of $C_7$ is $d = 4 - 2 + 1$ ( $2 =$ the number of points in $P$ on the side of $C_6$ of $e$ , $1 =$ the number of points in $P$ on the side $C_7$ of $e$ ). One possible path to compute simplicial depths of cells is shown in the figure, we can start for $C_1$ , which has simplicial depth $d = 0$ continue above step on $C_1, C_2, C_3, C_4, C_5, C_6, C_7, \dots$ path. . . .	24

4.1	The green and red sets have the same cardinal simplicial depth relative to the blue set. The green set's simplicial depth relative to the blue set is triple that of the red set's simplicial depth relative to the blue set by new Definition 4.1. These same examples demonstrate an analogous property for Tukey depth: the green set and the red set have the same generalized Tukey depth (by Definition 2.2) relative to the blue set, but their Tukey depths differ by new Definition 4.2. . . .	32
4.2	For a set $P$ of 5 points, and $P \cup Q$ in general position, placing a query point in each of the red cells gives the maximum achievable cardinal simplicial depth (with $CSD_P(Q) = 8$ out of 10 possible triangles) for a set $Q$ of 3 points, as in the figure on the left side. Placing three points anywhere in the blue cell gives the maximum simplicial depth according to Definition 4.1 (each point having a maximum of $SD_P(q) = 5$ ), as in the figure on the right side. . . . .	33
4.3	The set $P$ of 6 and 4 points are in the vertices hexagon and square, respectively. The sets made of two points that attain maximum cardinal closed simplicial depth are indicated in red (the two points should be placed close to the center and on the line connecting the center and two opposite vertices). In contrast, two overlapping points at the center of the regular $n$ -gon gives the maximum simplicial depth according to Definition 4.1. . . . .	34
4.4	The two images give an example for when the convex-hulls of $P$ and $Q$ are intersecting. In both examples, the convex Tukey depth of $Q$ relative to $P$ is determined by the point $x$ in $Q$ . Here, the blue line defines a halfspace which determines the convex Tukey depth of $Q$ . .	40
5.1	An example of sets $P$ and $Q$ in $\mathbb{R}^2$ , respectively, of 3 and 2 points, showing that property P.4.1 is not satisfied for the convex Tukey depth. In (A), points in $Q$ are interior to the triangle formed by the points in $P$ and, in (B), the two points in $Q$ are moving to infinity along fixed rays opposite to each other and parallel to bottom edge of the triangle. In both cases, the $CTD_P(Q)$ is 1. . . . .	48

6.1 Points in the primary plane  $(q, p_1 \dots p_n)$  are transformed to a set of lines  $(L_q, L_1 \dots L_n)$  in the dual plane. Then a new dual line  $L_{p'}$  is added to  $G$ . Finally, line intersection order in  $L_q$  is recalculated including the intersection between  $L_{p'}$  and  $L_q$  [27]. . . . . 54

# List of Tables

5.1	Comparing properties of simplicial depth and Tukey depth . . . . .	51
5.2	Comparing properties for depth measure of a set of points . . . . .	51

# Chapter 1

## Introduction

Depth measures quantify the centrality or eccentricity of an object relative to a set of objects. Compared to a distance metric, such as Euclidean distance, which can measure distance relative to a location estimator (or any point), depth measures provide a way to measure centrality relative to the whole population for any distribution. For univariate quantitative data, a natural definition for the depth of a point  $q$  in  $\mathbb{R}$  relative to a set  $P$  of points in  $\mathbb{R}$  is to measure how deeply nested  $q$  is in  $P$  by the lesser of the number of points of  $P$  less than  $q$ , and the number of point of  $P$  greater than  $q$ . This measure of depth provides a center-outward ordering of points in which outliers relative to  $P$  have low depth, whereas a median of  $P$  will be a deepest point. Various generalizations of this definition exist, allowing this univariate concept to be generalized to higher dimensions. Tukey [44] first introduced a notion of data depth, now referred to as Tukey depth or halfspace depth. The *Tukey depth* of a query point  $q$  relative to a set  $P$  of points is the minimum number of points of  $P$  in any closed half-space containing  $q$ :

**Definition 1.1 (Tukey depth of a query point [44]).** *Given a set  $P$  of  $n$  points in  $\mathbb{R}^d$  and a point  $q$  in  $\mathbb{R}^d$ , the Tukey depth (or half-space depth) of  $q$  relative to  $P$  is*

$$TD_P(q) = \min_{\substack{H \in \mathcal{H} \\ H \cap \{q\} \neq \emptyset}} |H \cap P|,$$

where  $\mathcal{H}$  is the set of all closed half-spaces in  $\mathbb{R}^d$ .

Based on Tukey depth, Donoho and Gasko [17] developed the idea of a multivariate median as a deepest point of the data set. Since then, several depth measures

have been introduced including convex hull peeling depth [7], Oja depth [34], regression depth [37] etc. with different properties and characteristics to characterize the depth of object of various types.

*Simplicial depth* introduced by Liu [29] is one such commonly use depth measure. The simplicial depth of a query point  $q$  relative to a set  $P$  of points is the number of simplices determined by points in  $P$  that contain  $q$ . Closed simplices were considered in the original definition of simplicial depth ( $q$  can lie on the simplex's boundary or in its interior). Liu [30] subsequently examined simplicial depth defined in terms of open simplices. In this thesis, we consider simplicial depth defined in terms of closed simplices and Tukey depth defined in terms of closed half-spaces.

**Definition 1.2 (Simplicial depth of a query point [29]).** *Given a set  $P$  of  $n$  points in  $\mathbb{R}^d$  and a point  $q$  in  $\mathbb{R}^d$ , the simplicial depth of  $q$  relative to  $P$  is*

$$SD_P(q) = \sum_{S \in \mathcal{S}} I(q \in S),$$

where  $\mathcal{S}$  denotes the set of  $\binom{n}{d+1}$  closed simplices, each of which is the convex hull of  $d + 1$  points from  $P$ , and  $I$  is the indicator function such that  $I(A) = 1$  if  $A$  is true and  $I(A) = 0$  otherwise.

Given  $q$  and  $P$  in  $\mathbb{R}^2$ , the simplicial and Tukey depths of  $q$  relative to  $P$  can both be computed in  $O(n \log n)$  time, where  $n = |P|$  [22, 38, 24]. Both also have matching worst-case lower bounds of  $\Omega(n \log n)$  time [4].

A median is a principal estimator defined by a depth measure that can summarize the location of the entire data set. A *median* of a set  $P$  is defined as any point of maximum depth relative to  $P$  for a given depth measure (note that the median isn't necessarily a point of  $P$ ). We refer to a *simplicial median* and *Tukey median*, respectively, for points of maximum simplicial depth and Tukey depth. Similarly, an *in-sample median* of  $P$  is a point of maximum depth among points in  $P$ . Given  $P$ , a simplicial median of  $P$  can be computed in  $O(n^4)$  time [5] and a Tukey median of  $P$  can be computed in  $O(n \log^3 n)$  time [26]. An in-sample simplicial median of  $P$  can be computed in  $O(n^2)$  time [22]. Similarly, in-sample Tukey depth can also be calculated in  $O(n^2)$  time [32].

Applications of data depth in Computer Science and Statistics are extensive

and still growing. In Computer Science, applications of data depth include pattern matching, geometric matching, shape fitting applications [2], clustering and classification [23]. In Statistics, many applications can be found in multivariate analysis [42], including robust location estimators (median and quantiles in higher dimensions), data visualization, regression analysis, hypothesis testing, multivariate control charts, as well as clustering and classifications. One of the key advantages of using data depth is that it leads to new nonparametric multivariate statistical analysis in which no distributional assumptions are needed.

Depth measures are typically defined to describe the location of a single query point (an individual) relative to a set of points (a population). In this thesis, we focus on the following two prominent problems related to data depth: (1) finding the depths of multiple query points relative to another set of points, and (2) deriving a single estimator for the depth of a set of query points relative to another set of points. These problems have recently gained interest (e.g., [35, 6, 8]). Pilz and Schnider [35] and Barba et al. [6] investigated the use of data depth of a set of query points to define quantiles in high-dimensional data sets. Barba et al. [6] defined the cardinal simplicial depth of a set  $Q$  as the number of simplices determined by points in  $P$  that contain at least one point of  $Q$ , and gave an  $O(N^{7/3} \log^{O(1)} N)$ -time algorithm in  $\mathbb{R}^2$ , where  $N = |P| + |Q|$ . Bertschinger et al. [8] defined the convex Tukey depth of a set  $Q$  relative to a set  $P$ . See Section 2.3.2 for more details.

## 1.1 Our Main Results

Defining and evaluating the depths of multiple query points and the depth of a query set of points have many applications. In some applications, we may need to query the depth of several points. Iteratively running an algorithm designed to calculate the depth of a single query point is inefficient, specifically when the size of the query point set is relatively large or some fraction of the querying point set. To resolve this, we will present in Chapter 3 several algorithms to calculate the depths of batched query points. The choice of which algorithm to apply to minimize the running time depends on the relative cardinality of the query point set to the input set.

In this thesis, we present algorithms for computing the depths of batched queries, and apply these to compute the depth of a query set of points. Given a set  $P$  of  $n$  points and a set  $Q$  of  $k$  points in  $\mathbb{R}^d$ , we introduce new definitions for the simplicial depth and Tukey depth of  $Q$  relative to  $P$ , along with algorithms for computing these in  $\mathbb{R}^2$ .

Specifically, We first consider three algorithms for computing the simplicial depths of batched query points in  $\mathbb{R}^2$ . These algorithms work for all values of  $n$  and  $k$ ; choosing which algorithm to apply to minimize running time depends on the relative values of  $k$  and  $n$ .

- Algorithm S.I applies a basic simplicial depth algorithm iteratively for each point in  $Q$  in  $O(kn \log n)$  time and  $O(n + k)$  space. This algorithm is suitable when  $|Q|$  is small relative to  $|P|$ , specifically, when  $k \in O(\frac{n}{\log n})$ .
- Algorithm S.II transforms points in  $Q$  and  $P$  to the dual plane to find the angular order of  $P$  around each point in  $Q$ , and then computes the simplicial depths of points of  $Q$  in  $O(n^2 + nk)$  time and  $O(n^2 + k)$  space. This algorithm is suitable when  $k \in \Omega(\frac{n}{\log n})$  and  $k \in O(n^3)$ .
- Algorithm S.III computes the simplicial depth inside each cell of a planar subdivision induced by  $P$  to compute the simplicial depths of points in  $Q$  in  $O(n^4 + k \log n)$  time and  $O(n^4 + k)$  space. This algorithm is suitable when  $|Q|$  is large relative to  $|P|$ , specifically, when  $k \in \Omega(n^3)$ .

Next, we present two algorithms for computing the Tukey depth of batched query points in  $\mathbb{R}^2$ . As above, these algorithms work for all values of  $n$  and  $k$ ; choosing which algorithm to apply to minimize running time depends on the relative values of  $k$  and  $n$ .

- Algorithm T.I applies a Tukey depth algorithm iteratively for each point in  $Q$  in  $O(kn \log n)$  time and  $O(n + k)$  space. This algorithm is suitable when  $|Q|$  is small relative to  $|P|$ , specifically, when  $k \in O(\frac{n}{\log n})$ .
- Algorithm T.II computes the Tukey depth contours of  $P$ , which determine a planar subdivision, for which the Tukey depth inside each cell is computed, and then computes the Tukey depths of points in  $Q$  in  $O(n^2 + k \log n)$  time

and  $O(n^2 + k)$  space. This algorithm is suitable when  $|Q|$  is large relative to  $|P|$ , specifically,  $k \in \Omega(\frac{n}{\log n})$ .

Then we discuss several applications of batched depth queries, including (1) finding the center outward ordering of a set of query points relative to the other set, (2) comparing multivariate distributions, and (3) constructing data visualization tools such as DD plots. A detailed discussion of these applications is in Chapter 3.

In Chapter 4, we focus on deriving a single estimator for the depth of a query point set relative to another set. We introduce definitions for the simplicial depth and Tukey depth of a set  $Q$  relative to a set  $P$ . These depth measures can be computed in  $\mathbb{R}^2$  by applying the algorithms above, leading to a computing time of  $O(\min\{kn \log n, n^2 + nk, n^4 + k \log n\})$  for simplicial depth, and in  $O(\min\{kn \log n, n^2 + k \log n\})$  for Tukey depth.

In Chapter 4, we then consider different statistical and probabilistic interpretation of the depth of a set query points based on the new definitions for simplicial and Tukey depth of sets. We show that in  $\mathbb{R}^d$ , the simplicial depth of  $Q$  relative to  $P$  is proportional to the expected number of points of  $Q$  contained in the simplex determined by  $d + 1$  points selected at random from  $P$ .

Applications of depth of a query set include (1) evaluating the centrality of a set of points relative to another set (e.g. the position of the players of a soccer team relative to the other team), (2) in classifying problems: a new set of points  $Q$  which comes from the same distribution (with similar characteristics) as one of given sets of points  $(P_1, \dots, P_i, \dots)$ . By comparing the depth of  $Q$  with each of the other sets  $P_i$ , we can classify  $Q$  as one of the sets  $P_i$  that gives the maximum depth for set  $Q$  relative to  $P_i$ , and (3) multivariate control charts. A detailed discussion of these applications is in Chapter 4.

Depth measures are commonly evaluated and compared in terms of their properties. We examine several common properties that are used to describe a depth of a single query. In Chapter 5, we extend and generalize these properties for the case of depth measures of sets. Further, we analyze these properties for the new definitions of simplicial depth and Tukey depth of a query set of points.

## 1.2 Thesis Outline

The remainder of this thesis is organized as follows. In Chapter 2, we give an overview of relevant literature about simplicial depth and Tukey depth. In Chapter 3, we present three algorithms for computing the simplicial depths of batched query points and two algorithms for computing the Tukey depths of batched query points in  $\mathbb{R}^2$ . Then, we discuss several application of depths of batched depth queries. In Chapter 4, we introduce new definitions for the simplicial depth and Tukey depth of a set  $Q$  relative to set a  $P$ . Further, in this chapter, we describe how to apply our algorithm in Chapter 3 to compute the simplicial depth and Tukey depth of a query set and present different statistical perspectives for our definitions of simplicial and Tukey depth of a query set. With that, we propose several applications of the depth of set. In Chapter 5, we extend and generalize existing data depth properties for depth measures of a set of points. Further, we analyze these properties for the new definitions of simplicial depth and Tukey depth for set and present a comparison with other definitions of depth of a set of points. Finally, Chapter 6 summarizes the main finding of the thesis and concludes the thesis suggesting possible avenues for future research.

# Chapter 2

## Literature Review

To set the stage for new results obtained about the depth of multiple query points, in this chapter, we first discuss previous work related to data depth in Section 2.1, depth medians in Section 2.2, and existing literature on the depth of a query set of points and related problems in Section 2.3.

### 2.1 Data Depth

Consider the following problem: you are given a set  $P$  of points and a query point  $q$ ; the task is to measure the location/centrality (provide a location/centrality estimator) of the query point  $q$  relative to the given set  $P$  of points. For instance, in a univariate data set, the location of a query point can be described by the distance relative to one of the location estimators for the data set (median or mean) or by the rank of the query point within the set of other points. Solving this problem gets more complicated when more variables are available to describe data points, that is, when the dimension of data increases.

Popular distance measures cannot directly adapt to measure the centrality of the query point relative to the other set because these distance measures (like Euclidean distance) are calculated relative to a location estimator. The use of location estimators is to summarize the location of the whole data set. But, measuring the depth of a query point requires estimating the centrality relative to the entire data set. Therefore, the depth of a query relative to the whole data set is not summarized by distance to a location estimator; additionally, these location estimators could

suffer from the issue of robustness. Furthermore, Euclidean distance can only be applied to uncorrelated variables. In comparison, Mahalanobis distance can measure the distance of a query point within correlated data. Typically, Mahalanobis distance measures the distance relative to the mean, accounting for the covariance matrix of variables. This popular outlyingness measure is most appropriate under a model assumption of elliptical symmetric probability distribution of the underlying population.

Therefore, several depth statistics were introduced to quantify centrality, each potentially having different properties (see Chapter 5) and characteristics. This includes efficient computability, robustness against outliers, and the ability to measure centrality in a distribution-free (nonparametric) way, that is, without relying on any distribution feature on the underlying population. This section discusses two common depth measures: Tukey depth and simplicial depth.

### 2.1.1 Tukey Depth

The first notion of data depth known as halfspace depth, location depth or Tukey depth was introduced by Tukey in 1975 [44]. In that work, Tukey proposed several analysis tools for bivariate data sets based on halfspace depth. First, he defined the Tukey depth of a query point  $q$  relative to a particular point set  $P$  in  $\mathbb{R}^2$ , denoted here as  $TD_P(q)$  as the minimum number of points in any closed halfspace containing  $q$  (Definition 1.1). Second, he defined the multivariate Tukey median as a generalization to the univariate median, to be a deepest point of the data set, that is a point of maximum Tukey depth. Then he presented the idea of depth contours as a graphical tool for visualizing bivariate data sets. Tukey depth has become one of the most famous tools in multivariate data analysis because of its fast computability and because of its properties (see Section 5 for more details). Applications of Tukey depth in multivariate analysis include the calculation of location estimators like medians and quantiles in higher dimensions, defining of outlyingness functions, visualization tool for data, principal component analysis, classification [33, 42], etc. Because of the scale and variety of applications, designing efficient algorithms for computing data depth and solving related problems has attracted a great deal of attention from researchers in the field of Computational Geometry.

In  $\mathbb{R}^2$ , Aloupis et al. [4] showed that computing the Tukey depth of an arbitrary query point relative to a set of  $n$  requires  $\Omega(n \log n)$  time in the worst case. Matching this lower bound, in  $\mathbb{R}^2$ , the Tukey depth of a query point can be computed in  $O(n \log n)$  time relative to a set of  $n$  points [38]. Different approaches to computing the Tukey depth were studied, including a Monte Carlo approximation algorithm by Chen et al. [15] and an output-sensitive algorithm by Bremner et al. [11]. Rousseeuw et al. [40] introduced an algorithm to compute the Tukey depth of a given point in  $\mathbb{R}^d$  for  $d > 2$  in  $O(n^{(d-1)} \log n)$  time. For higher dimensions ( $\mathbb{R}^d$ ), the maximum Tukey depth is between  $\lceil n/(d+1) \rceil$  and  $\lceil n/2 \rceil$  [13].

### 2.1.2 Simplicial Depth

This section presents an overview of previous work related to computing the simplicial depth for a single query point. The notion of simplicial depth was introduced by Liu [29]. In  $d$ -dimensional space ( $\mathbb{R}^d$ ), a simplex is the convex hull of a set of  $d+1$  points. The simplicial depth ( $SD_P(q)$ ) of a query point  $q$  with respect to set  $P$  is defined as the number of simplices whose vertices are points in  $P$ , that contain the query point  $q$  (Definition 1.2). In Statistics, simplicial depth is also studied by normalizing with the total number of simplices, leading to a probability distribution. Further, two major versions exist for the simplicial depth definition: one definition uses closed simplices, and the other open simplices.

Liu considered closed simplices in her first definition of simplicial depth [29]. In this case, if a query point lies on a boundary of a simplex, it is considered to be included in the simplex. Therefore, according to this definition, simplices which have  $q$  lying on their boundary count towards the depth of  $q$ . Later, Liu et al. defined simplicial depth based on open simplices. In this definition, simplices which have  $q$  lying on their boundary do not count towards the depth of  $q$  [30]. However, both of these definitions cause irregularities of depth values at boundaries of simplices. A revised definition was proposed to correct these irregularities by taking the average of depth values given by the above two definitions [12]. Observe that the simplicial depth of a point  $q$  relative to a set  $P$  of points is identical under either definition when  $P \cup \{q\}$  is in general position. On the other hand, the location and depth of a deepest point differs between the two definitions (see Section 2.2).

In two-dimensional space, a simplex is a triangle. Therefore, in  $\mathbb{R}^2$  the simplicial depth of a point  $q$  is the number of triangles, whose vertices are points of  $P$  that contain  $q$ . Even before Liu's definition for simplicial depth, an algorithm to count the number of triangles containing a query point in  $O(n \log n)$  time was presented [24]. Similar algorithms were studied in other independent works by Gil et al. [22] and Rousseeuw and Ruth [38]. Each algorithm involves sorting the set  $P$  around the query point  $q$  (radial ordering). This step takes  $O(n \log n)$  time and dominates the time complexity. Once the radial order of the points of  $P$  is determined, each of these algorithms computes the simplicial depth of  $q$  in  $O(n)$  time [22]. Further, Aloupis et al. [4] proved a lower bound of  $\Omega(n \log n)$  time to compute the simplicial depth in  $\mathbb{R}^2$ .

When  $q$  and the points in  $P$  are in  $d$  dimensional space, a straightforward method to calculate the simplicial depth of a query point takes  $O(n^{d+1})$  time [16]. It iteratively checks whether a given query point is contained in each of the simplices. For  $\mathbb{R}^3$ , Rousseeuw and Ruts were the first to present an approach that adapts the technique used to compute simplicial depth in  $\mathbb{R}^2$  for computing simplicial depth in  $\mathbb{R}^3$  [38]. An algorithm for simplicial depth in  $\mathbb{R}^3$  with  $O(n^2)$  time was first presented by Gil et al. [22]. Later, Cheng and Ouyang [16] clarified the method presented in [38] and showed a flaw in the algorithm proposed by Gil et al. Further, they introduced a modified version of the algorithm by correcting a this flaw in [22] and stated that the algorithm has  $O(n^2)$  time complexity. Moreover, Cheng and Ouyang [16] discussed a method to adapt the technique used to compute simplicial depth in  $\mathbb{R}^3$  to compute simplicial depth in  $\mathbb{R}^4$  in  $O(n^4)$  time. In 2016, Afshani et al. [1] proposed methods to compute simplicial depth in  $O(n^d \log n)$  time for  $d > 4$ . An algorithm presented by Pilz et al. [36] further improved this time bound to  $O(n^{d-1})$ .

## 2.2 Depth-Related Medians

Location estimators such as median, mean, and mode are used to summarize the location of a whole data set. The robustness of those location estimators is important to evaluate how much an estimator changes with perturbations of data. Among those, the mean is a non-robust estimator because taking one point far outside the

data set also moves the mean arbitrarily far outside the rest of the data set. A median, on the other hand, is a robust estimator, as moving a point or some small fraction of the points to infinity does not change its location. Because of this, when the probability distribution underlying a data set is unknown, a median is often preferred as a location estimator. The fraction of points of a data set that can be moved to infinity without making the location estimator also move to infinity, known as the breakdown point, can be used to measure the robustness of a median [31]. In one dimensional space, the median has a breakdown point of  $\lfloor \frac{n-1}{2} \rfloor \frac{1}{n}$ ; that is, no more than  $\lfloor \frac{n-1}{2} \rfloor$  points of a set of  $n$  points can be moved with the median not moving outside the points that remain unchanged. In [31], Rousseuw and Leeuwen discussed breakdown points of several estimators. A survey paper by Aloupis [3] summarized several depth measurements and medians.

### 2.2.1 Tukey Depth Median and Contours

For a set of points in  $\mathbb{R}^d$ , a Tukey median is a point with maximum Tukey depth. It is a robust and invariant estimator under affine transformation [38]. For a set  $P$  of  $n$  points in general position in  $\mathbb{R}^d$ , the maximum Tukey depth lies between  $\lceil n/(d+1) \rceil$  and  $\lceil n/2 \rceil$  [17]. A Tukey median in  $\mathbb{R}^2$  can be found in  $O(n \log^3 n)$  time [26].

Tukey depth contours are a collection of nested polygons that partition the plane into regions of equal Tukey depth. Tukey [44], suggested using the corresponding depth contours to visualize bivariate data. As shown by Miller et al. [32], Tukey depth contours can be computed for a given set  $P$  in  $O(n^2)$  time and space [32]. To do so, their algorithm maps  $P$  to the dual plane, forming a line arrangement, over which a topological sweep identifies contour lines and intersection points in the primal plane.

### 2.2.2 Simplicial Median

In  $\mathbb{R}^2$ , multiple authors showed that the maximum achievable simplicial depth for any point  $q$  is  $\frac{n^3-4n}{24}$  when  $n$  is even, and  $\frac{n^3-n}{24}$ , when  $n$  is odd when all the points are in general position [10]. Further, Boros and Furefi [10] showed that there is a query point of depth at least  $\frac{n^3}{27} + O(n^2)$  for any set  $P$  of  $n$  points in general position.

Gil et al. [22] showed that the simplicial depth of all points of  $P$  can be computed

in  $O(n^2)$  time in  $\mathbb{R}^2$ , so the in-sample simplicial median can be computed in  $O(n^2)$  time. Given a set  $P = \{p_1, \dots, p_n\}$  of points in  $\mathbb{R}^2$ , Lee and Ching [27] showed that the radial order of  $P \setminus \{p_i\}$  with respect to  $p_i$  for all  $i \in \{1, \dots, n\}$  can be determined in  $O(n^2)$  time. To accomplish this, they consider the set  $L = \{l_1, \dots, l_n\}$  of lines, where  $l_i$  is the dual line to the point  $p_i \in P$ . Lee and Ching established a linear relationship between the angular order of  $P \setminus \{p_i\}$  around  $p_i$  and the order of intersections between  $l_i$  and  $L \setminus \{l_i\}$ . Using this technique and traversing the planar graph determined by  $L$ , the sorted list of points in  $P \setminus \{p_i\}$  with respect to  $p_i$  can be constructed in  $O(n)$  time. Construction of the planar graph determined by the set  $L$  of  $n$  lines takes  $O(n^2)$  time and  $O(n^2)$  space [14]. Therefore, the radial order of  $P \setminus \{p_i\}$  with respect to  $p_i$ , for all  $i \in \{1, \dots, n\}$ , can be found in  $O(n^2)$  time and  $O(n^2)$  space. Given the order of points of  $P$  around  $p_i$ , the depth of  $p_i$  can be computed in  $O(n)$  time. Therefore, using the sorting step in [27], the simplicial depth of all  $n$  points in  $P$  can be obtained in  $O(n^2)$  time [22]. Consequently, an in-sample simplicial median can be computed in  $O(n^2)$  time [22]. An algorithm solving a similar problem was studied independently by Khuller and Mitchell [24].

In the case where simplicial depth is defined based on closed simplices, a median lies at an intersection of simplex boundaries [5]. Each boundary line segment joins a pair of points in  $P$ . Rousseeuw and Ruts [38] introduced a method to find a simplicial median using intersection points in  $O(n^5 \log n)$  time. Aloupis et al. [5] derived a faster algorithm to compute a simplicial median in  $O(n^4 \log n)$  time and  $O(n^2)$  space. By applying a topological sweep technique [19], Aloupis et al. further improved the running time to find a simplicial median in  $O(n^4)$  time and  $O(n^2)$  space [5].

The simplicial median algorithm of Aloupis et al. [5] begins by computing the simplicial depths of all points in  $P$ . This step takes  $O(n^2)$  time, as described above. Then, it computes the intersection points of each segment sorted along the  $x$ -axis, taking  $O(n^2 \log n)$  time per segment. Next, the algorithm determines which endpoint of each segment has maximum depth, and the depth of its adjacent intersection (endpoints of these line segments are points in  $P$ ) which takes  $O(n \log n)$  time per segment. Finally, the simplicial depth is calculated at all intersections along a line segment in  $O(n^2)$  time; we apply a similar technique in our algorithm described in

Section 3.1. To execute these steps for  $O(n^2)$  line segments takes  $O(n^4 \log n)$  time and  $O(n^2)$  space. While the algorithm calculates the simplicial depth at intersection points, it maintains the intersection point with maximum depth to find a simplicial median. As mentioned above, Aloupis et al. [5] also proposed an alternative algorithm using topological sweeping. In this case, instead of processing each line segment sequentially, the algorithm calculates the simplicial depth at intersections while sweeping the arrangement of line segments. Using this approach, a simplicial median can be computed in  $O(n^4)$  time and  $O(n^2)$  space.

## 2.3 Depths of Multiple Query Points

There are various ways of defining the depth of a query point  $q \in \mathbb{R}^d$  relative to a point set  $P \subseteq \mathbb{R}^d$ . Two of the most prominent and well-studied depth measures are simplicial depth  $SD_P(q)$  and Tukey depth  $TD_P(q)$ , as we discussed in previous sections.

Suppose that instead of a single query point  $q$ , we are given a set  $Q$  of  $k$  query points. Two interrelated questions then arise.

1. How fast can we compute the simplicial depth  $SD_P(q)$  or the Tukey depth  $TD_P(q)$ , for every  $q \in Q$ ? We will refer to this as the problem of computing the depths of batched query points or, simply, batched depth queries.
2. Can one define generalized simplicial and Tukey depths, for the whole set  $Q$  relative to  $P$ ? We will refer to this as the problem of computing the depth of a query point set.

The problem of defining the depth of a query set has recently gained interest (e.g., [35, 6, 8]). Barba et al. [6] also investigated the use of a data depth measure for a set of points to define quantiles in high-dimensional data sets, by generalizing simplicial depth for a set. Pilz and Schnider [35] extend the concept of central point to that of a central set and use this to define quantiles for high-dimensional data set through a generalization of Tukey depth. Another recent study by Bertschinger et al. [8] presented several different definitions for the Tukey depth of a set of query points. In what follows, we elaborate on these findings.

### 2.3.1 Simplicial Depth of a Query Set

Recently, Barba et al. [6] introduced a definition for the simplicial depth of multiple query points as follows. To disambiguate between Definitions 2.1 and our definition for simplicial depth of a set (Definition 4.1), we refer to Definition 2.1 as the *cardinal simplicial depth* because it corresponds to the cardinality of the set of non-empty simplices.

**Definition 2.1 (Cardinal simplicial depth of multiple query points [6]).**

The cardinal simplicial depth of a set  $Q \subseteq \mathbb{R}^d$  with respect to a set  $P \subseteq \mathbb{R}^d$  is

$$CSD_P(Q) = \sum_{S \in \mathcal{S}} I(Q \cap S \neq \emptyset), \quad (2.1)$$

where  $\mathcal{S}$  denotes the set of  $\binom{n}{d+1}$  open simplices, each of which is the convex hull of  $d+1$  points from  $P$ , and  $I$  is an indicator function such that  $I(A) = 1$  if  $A$  is true and  $I(A) = 0$  otherwise.

In other words, cardinal simplicial depth counts the number of open simplices with  $d+1$  vertices in  $P$  that contain at least one point  $q \in Q$ . In  $\mathbb{R}^2$ , it counts how many open triangles formed by points in  $P$  contain at least one point of  $Q$ .

As in the work of Pilz et al [35], Barba et al. [6] also aimed to find a high-dimensional analogue to quantiles. In  $\mathbb{R}^2$ , the cardinal simplicial depth of two points can be computed in  $O(n \log n)$  time. When  $P$  and two query points are in general position, the maximum cardinal simplicial depth is  $\frac{n^3}{12} + \frac{n^2}{3}$ , whereas the maximum achievable simplicial depth for a single query point is  $\frac{n^3}{24} - \frac{n}{6}$ . In the special case when  $P$  (in  $\mathbb{R}^2$ ) is in convex position, a set of  $k$  (in  $\mathbb{R}^2$ ) query points has an upper bound of  $\frac{n^3 k}{6(k+3)} + \frac{3n^2}{2(k+3)} - \frac{5n}{24}$  for its cardinal simplicial depth [6].

From the above definition, we can observe that when more query points are available, the cardinal simplicial depth of  $Q$  will be maximized by spreading out its points within the convex cloud of points of  $P$  to capture as many simplices formed by points in  $P$  as possible. In  $\mathbb{R}^2$ , when both  $P$  and  $Q$  are in general positions, Barba et al. [6] presented an algorithm to compute the cardinal simplicial depth of  $Q$  in  $O(N^{7/3} \log^{O(1)} N)$  time, where  $N = n + k$ .

### 2.3.2 Tukey Depth of a Query Set

Recently, Pilz and Schnider [35] studied the extension of quantiles to a high-dimensional data. The objective was to find a small set of points that could be used to summarize or represent a bigger set of points. For that, they presented the following generalized Tukey depth for a set of query points and defined quantiles as the set with maximum generalized Tukey depth among all sets of a given fixed size [35].

**Definition 2.2 (Generalized Tukey depth for multiple points [35]).** *The generalized Tukey depth of a set  $Q \subseteq \mathbb{R}^d$  with respect to a set  $P \subseteq \mathbb{R}^d$  is*

$$GTD_P(Q) = \min_{\substack{H \in \mathcal{H} \\ Q \cap H \neq \emptyset}} \frac{|H \cap P|}{|H \cap Q|},$$

where  $\mathcal{H}$  is the set of all closed half-spaces in  $\mathbb{R}^d$ .

We emphasize here that the above definition is appropriate for sets. Using the point of view of the current study, this problem is not one of multiple query points, but rather, it is one of computing the depth of a query point set. Indeed, the above definition does not entail calculating a depth for each query point of  $Q$ , which we term batched queries.

In  $\mathbb{R}^d$ , Pilz and Schnider proved that there is always a set  $Q$  of  $k$  points that has a generalized Tukey depth of at least  $\frac{1}{kd+1}$ . Then in  $\mathbb{R}^2$ , they derived an  $O(n \log^3 n)$  time algorithm for finding two points  $p_1$  and  $p_2$ , such that any halfplane containing one of  $p_1$  and  $p_2$  contains at least  $n/5$  points of  $P$ , and each closed halfspace containing both  $p_1$  and  $p_2$  contains at least  $2n/5$  points of  $P$ .

A recent study by Bertschinger et al. [8] defined two variations of Tukey depth for a set  $Q$  relative to  $P$ , affine Tukey depth and convex Tukey depth as follows.

**Definition 2.3 (Affine Tukey depth of a set of query points [8]).** *Given a set  $P$  of  $n$  points in  $\mathbb{R}^d$  and a set  $Q$  of  $k + 1$  points in  $\mathbb{R}^d$ ,  $k < d$ , which spans a unique  $k$ -flat  $F$ , the Affine Tukey depth of  $Q$  ( $ATD_P(Q)$ ) is the minimum number of points of  $P$  in any closed halfspace containing  $F$ .*

**Definition 2.4 (Convex Tukey depth of a set of query points [8]).** *Given a set  $P$  of  $n$  points in  $\mathbb{R}^d$  and a set  $Q$  of  $k$  points in  $\mathbb{R}^d$ , the Convex Tukey depth*

of  $Q$  ( $CTD_P(Q)$ ) is the minimum number of points of  $P$  in any closed halfspace containing  $Q$ .

### 2.3.3 Depth Histograms

We discussed in Section 2.1 that depth measures differ from the distance between the query point  $q$  and a location estimator of  $P$ . Further, the exact location of  $q$  is also sometimes insignificant as many depth measures assign equal depth within cells bounded by lines/hyperplanes connecting points of  $P$ . In fact, many of popular depth measures such as Tukey depth, simplicial depth, and convex hull peeling depth depend on the combinatorial structure of  $P$ . For instance, Tukey depth and simplicial depth can be computed using the angular order of points in  $P$  around  $q$ . One important deviation from this is *Oja depth* which is defined as the sum of the area of simplices formed by points in  $P$  that contain  $q$ . In what follows, we briefly summarize algorithms for computing the Tukey depth and simplicial depth of a query point in the plane discussed by Aloupis et al. [4].

#### Algorithms for Simplicial Depth and Tukey Depth for a Single Query Point in the Plane

Consider a set  $P$  of  $n$  points and a query point  $q$  in  $\mathbb{R}^2$ . Aloupis et al. [4] presented the following algorithm for computing the simplicial and the Tukey depth of  $q$  relative to  $P$ . These two algorithms require to first sort the set  $P$  radially around  $q$ . This step takes  $O(n \log n)$  time. Next steps of the algorithms take only  $O(n)$  time. To find the Tukey depth of  $q$ , we find the number of points in  $P$  in each halfspace which contains  $q$ . To find the simplicial depth of  $q$ , we can compute the number of triangles that do not contain  $q$  and subtract it from the total number of triangles formed by selecting three points of  $P$ . Interestingly, we have that if three points lie in one half space through  $q$  and  $p_i$ , then that triangle formed by these three points does not contain  $q$ , see Figure 2.1. As a result, the triangles that do not contain  $q$  can be found by considering all halfspaces through  $q$  and the associated triangles constructed from points in these halfspaces.

To compute the Tukey or simplicial depths, we first sort the set  $P$  radially around  $q$ . Then, we define a directed line  $L$  from  $q$  to any point  $p_i$ . Next, extend  $L$  on the

other side of  $q$ ; this extension is defined as  $S$ . Further, combining  $L$  and  $S$  into one line (denoted  $LS$ ) going through the point  $q$  creates two halfspaces. Let  $h_i$  be the number of points on the left halfspace of the  $LS$  line. When calculating  $h_i$ , any point on  $L$  is included and any point on  $S$  is excluded. Since the points are sorted around  $q$ , we can find  $h_i$  in  $\theta(1)$  time. To find the next halfspace, we can rotate the line  $LS$  around  $q$  up to the next point  $p_{i+1}$ , and then find the next halfspace count  $h_{i+1}$ . For Tukey depth, keep track of the minimum  $h_i$  and report it as the Tukey depth of  $q$ , this is  $TD_P(q) = \min\{h_i\}$  for  $i \in \{1 \dots n\}$ . For simplicial depth, by selecting  $p_i$  and two other point from  $h_i$ , we can count the number of triangles that do not contain  $q$  as  $\binom{h_i}{2}$ . Then rotate line  $LS$  around  $q$ , and redo the calculation for the next halfspace. This process ends when  $LS$  has rotated a full rotation. The simplicial depth of  $q$  is  $SD_P(q) = \binom{n}{3} - \sum_{i=1}^n \binom{h_i}{2}$ .

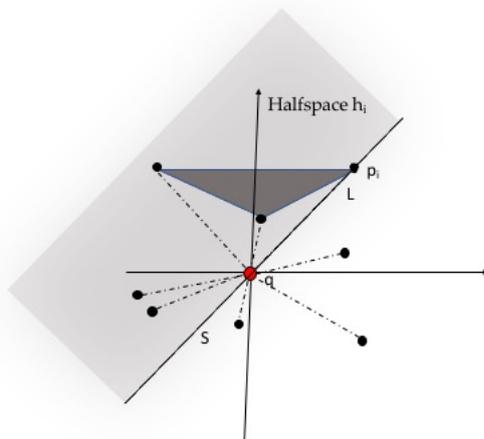


Figure 2.1: Simplicial depth calculation

As described above, both calculations start counting the number of points to the left of the line  $LS$ . Now, we note the values of  $h_i$ 's can be plotted in a frequency chart which tracks the number of halfspaces with  $j$  points on the left side of the line  $LS$ , for  $j \in \{0, \dots, n\}$ . This distribution was discussed in several studies and is referred to as a  $\psi$ -histogram [18],  $l$ -vector [8], or frequency vector [36]. Durocher et al. [18] first introduced the idea of using  $\psi$ -histograms as a characterization of the combinatorial structure of the point set  $P$ . They showed that the  $\psi$ -histogram contains sufficient information to compute several depth measures for  $q$  with respect to  $P$ , including Tukey depth, simplicial depth, perihedral depth and Eutomic depth.

The study by Bertschinger et al. [8], which defined the Affine Tukey depth and convex Tukey depth of a set of query points, discussed Tukey depth histograms of  $k$ -flats. A depth histogram summarizes how many points in a point set have the same depth for each possible depth value in the range of possible depths.

## Chapter 3

# Computing the Depths of Batched Query Points

In this chapter, we consider solving the following question: suppose that instead of a single query point  $q$ , we are given a set  $Q$  of  $k$  batched query points; how fast can we compute the simplicial depth  $SD_P(q)$ , respectively, the Tukey depth  $TD_P(q)$ , for every  $q \in Q$ ? For a given query point and  $P$  in  $\mathbb{R}^2$ , the simplicial and Tukey depths of  $q$  relative to  $P$  can both be computed in  $O(n \log n)$  time, where  $n = |P|$ . Algorithms have been provided and studied in [22, 38, 24]. When the size of  $Q$  is relatively small compared to  $P$ , iteratively applying the above algorithms to each point in  $Q$  is efficient. But when the size of  $Q$  is large or when its size is some fraction of  $P$ , this method can be inefficient.

We study this problem and describe algorithms that compute the simplicial depths and Tukey depths for  $k$  points in a set  $Q$  relative to a set  $P$  of  $n$  points, where  $P \cup Q$  is in general position in  $\mathbb{R}^2$ . For both depth measures, which algorithm is fastest depends on the relative values of  $k$  and  $n$ . For simplicial depth, we present three algorithms: Algorithm S.I is not new [22, 38, 24]; Algorithms S.II and Algorithm S.III are new. For Tukey depth we propose two algorithms: Algorithms T.I and Algorithm T.II are based on existing algorithms for Tukey depth and Tukey depth contours.

## 3.1 Algorithms for Computing Simplicial Depths of Batched Query Points

### 3.1.1 Algorithm S.I

Section 2.3.3 mentioned an algorithm for computing the simplicial depth of single query point  $q$  relative to a set  $P$  of  $n$  points in  $\mathbb{R}^2$  in  $O(n \log n)$  time. Similarly, three algorithms in [22, 38, 24] also compute the sorted angular order of  $P$  around  $q$  as the first step, which dominates the time complexity of the rest of the algorithms. Therefore, resulting algorithms also have the same time bound of  $O(n \log n)$ .

To find the batched depth of points in  $Q$  relative to  $P$ , when the number of query points  $k$  is small relative to  $n$ , a straightforward approach for computing the depths of  $k$  points is to iteratively compute the simplicial depth of each query point using one of these existing algorithms. This takes  $O(kn \log n)$  time and  $O(n + k)$  space to store the angular order of  $P$  around a query point (this space is reused for each query point) and store the depth of  $k$  query points. There is a lower bound of  $\Omega(n \log n)$  on the worst-case time required for computing the simplicial depth of a single point [4]. Consequently, this approach is optimal when  $k \in O(1)$ . This gives the following lemma:

**Lemma 3.1.** *Given a set  $P$  of  $n$  points and a set  $Q$  of  $k$  query points in general position in  $\mathbb{R}^2$ , Algorithm S.I computes  $SD_P(q)$  for every  $q \in Q$  relative to  $P$  in  $O(kn \log n)$  time and  $O(n + k)$  space.*

### 3.1.2 Algorithm S.II

Algorithm S.I is efficient when  $k$  is small relative to  $n$ , but more efficient approaches are possible for larger values of  $k$ . We describe an algorithm that computes the simplicial depth of points in  $Q$  relative to  $P$  in  $O(n^2 + nk)$  time and  $O(n^2)$  space. We use an approach similar to the in-sample simplicial median algorithm described by Gil et al. [22] and the triangle counting problem examined by Khuller and Mitchell [24]. This problem is the same as the in-sample simplicial median in  $\mathbb{R}^2$ , which compute for every point  $p \in P$  the number of triangles (formed by points in  $P$ ) that contain it.

Our algorithm has two steps: (1) compute the sorted radial order of the  $n$  points of  $P$  around in  $O(n)$  time for each point of  $Q$ . This step requires additional  $O(n^2)$  preprocessing time. (2) use the ordering to compute the simplicial depth of each point in  $Q$ . As described in simplicial depth calculation algorithm in Section 2.3.3, once the sorted order of set  $P$  around a point  $q$  is known, the simplicial depth of  $q$  can be calculated in  $O(n)$  time [22]; consequently, the second step can be executed in  $O(kn)$  time to compute simplicial depth of all points in  $Q$ .

To perform Step 1, we modify the method described by Gil et al. [22] and Khuller and Mitchell [24]. First, the sets  $P$  of  $n$  points ( $P = p_1, \dots, p_n$ ) and  $Q$  are transformed into line sets  $L_P$  and  $L_Q$  in the dual plane, respectively. The sorted order of  $P$  around a point  $q$  can be obtained by considering the intersection order of  $L_P$  with the dual-line  $L_q$  using a method described by Lee et al. [27]. As shown in Figure 3.1, the angular order of  $P$  around  $q$  can be computed using intersection order of the dual line set  $L_P$  with dual the line  $L_q$ . This is based on the principle that when a line on  $q$  rotates around  $q$  in the clockwise direction, its dual point moves from plus infinity ( $+\infty$ ) to minus infinity ( $-\infty$ ) along  $L_q$ . According to Lee's method, to find the angular order,  $P$  is partitioned into four quadrants based on the origin  $O$  and  $q$  (using the  $Oq$  line) at  $q$ . For each quadrant, maintain a separate list to insert points as was done finding the intersection order of dual lines in  $L_P$  along  $L_q$ . Then, identify the quadrant of each  $p_i$  and append  $p_i$  to the list in the relevant quadrant. This step requires  $O(n)$  time for each point in  $Q$ .

The planar graph construction method in [14] can be implemented to find the intersection order of lines in the  $L_P$  set with each line  $L_q \in L_Q$ . For Step 1, we construct a graph  $G$  of the arrangement of lines induced by  $L_P$  incrementally by introducing one line at a time, and construct the doubly connected edge list of  $L_P$ , which requires  $O(n^2)$  time and  $O(n^2)$  space as a preprocessing. Then, we continue this process by temporarily adding each line in  $L_q$  to  $G$ , and finding the order of intersections of lines in  $L_P$  with  $L_q$  by traversing the sequence of edges in  $G$  along  $L_q$ , which takes  $O(n)$  time. While finding the intersections, the sorted angular order of  $P$  around each point  $q$  is obtained using the method described above in  $O(n)$  time. The whole Step 1 requires preprocessing to compute and store  $G$  in  $O(n^2)$  time and  $O(n^2)$  space, and to find the sorted angular order of  $P$  around each query

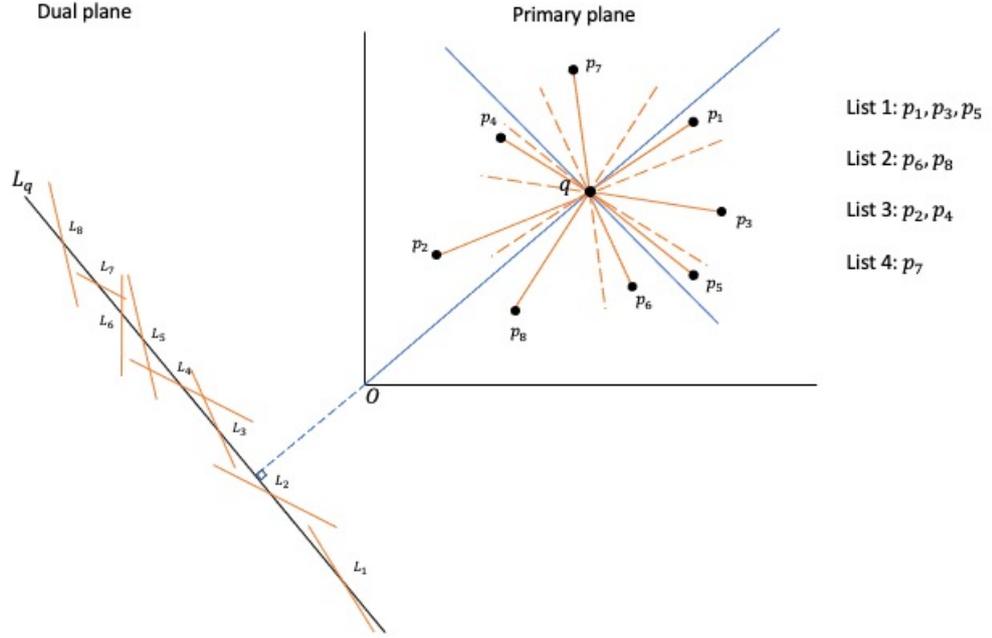


Figure 3.1: Points in the primary plane ( $q, p_1 \dots p_8$ ) are transformed to the set of lines ( $L_q, L_1 \dots$ ) in the dual plane. Here  $O$  is the origin and the line connecting origin and  $q$  ( $Oq$  line) partitions the primal plane into four quadrants at  $q$ . While finding the intersection order of  $L_P$  lines in  $L_q$ , insert  $p_i$  to the list of the relevant quadrant. Combining four lists gives the angular order of  $P$  around  $q$ . This figure replicates an example illustrated by Lee and Ching [27, Figure 3].

point in  $O(n)$  time. In Step 2, the simplicial depth of each point  $q \in Q$  relative to  $P$  can be found in  $O(n)$  time using the angular order of points of  $P$  around  $q$  obtained in Step 1. This takes  $O(nk)$  time for all points in  $Q$ , giving a total time of  $O(n^2 + nk)$  and  $O(k)$  space to store depth values.

Step 1 requires finding the order of intersections between  $L_P$  and each line in  $L_q$ . Finding the order of intersections between one line and a set of  $m$  lines can be achieved using one of various methods: (a) incremental planar graph construction in  $O(m^2)$  time and  $O(m^2)$  space [14], (b) line sweeping in  $O(m^2 \log m)$  time [43, 20], or (c) topological sweeping in  $O(m^2)$  time and  $O(m)$  space [19]. Despite its lower costs as a function of  $m$ , when applied to our problem, topological sweeping takes  $O(n^2 + k^2)$  time and  $O(n + k)$  space because it processes additional intersections in  $L_P$  and  $L_Q$  that are not needed for Step 1. The most efficient method for finding the ordered intersections between  $L_P$  and each  $L_q$  line is the incremental planar graph construction, which leads to an overall  $O(n^2 + nk)$  time and  $O(n^2)$  space. This gives the following lemma:

**Lemma 3.2.** *Given a set  $P$  of  $n$  points and a set  $Q$  of  $k$  query points in general position in  $\mathbb{R}^2$ , Algorithm S.II computes  $SD_P(q)$  for every  $q \in Q$  relative to  $P$  in  $O(n^2 + nk)$  time and  $O(n^2 + k)$  space.*

### 3.1.3 Algorithm S.III

When  $k$  is large relative to  $n$ , construct the arrangement  $L$  formed by lines between every pair of points in  $P$ . This arrangement partitions the plane into  $\Theta(n^4)$  convex cells, when  $P$  is in general position. By modifying the simplicial median calculation algorithm, we can compute the depth of all cells in  $O(n^4)$  time. In Section 2.2.2, we described an algorithm for finding a simplicial median in  $O(n^4)$  time and  $O(n^2)$  space introduced by Aloupis et al. [5]. They consider the arrangement of line segments connecting every pair of points in  $P$ , which also has  $O(n^4)$  intersections and  $O(n^4)$  cells. This method computes the number of points on each side of each line segment of  $P$  in  $O(n^3)$  time. Further, they showed that starting from a known depth value on a line segment, by processing each intersection point in  $O(1)$  time, the simplicial depth along the line segment can be computed in  $O(n^2)$  time [5]. We adapt this depth-finding method along a line segment to find the simplicial depth of cells in our arrangement  $L$  as described below.

Each line  $l$  in  $L$  is first partitioned into three sections based on the two points  $p_1$  and  $p_2$  in  $P$  that determine  $l$ : the line segment between  $p_1$  and  $p_2$  (color this segment blue) and two rays (color the rays red) on  $l$  rooted respectively at  $p_1$  and  $p_2$ . In the arrangement determined by  $L$ , only the blue segments are boundaries of simplices. Therefore, when crossing from one cell to an adjacent cell, the depth potentially changes if the two cells share a blue line segment on their common boundary. Similarly, if the two cells share a red segment on their common boundary, then both cells have the same simplicial depth.

We compute the number of points on each side of each line in  $L$  in  $O(n^3)$  time. The algorithm traverses the arrangement and computes the simplicial depth in adjacent cells, starting from a cell  $C_i$  with known simplicial depth. To find the simplicial depth of a cell  $C_j$  that shares a blue edge with  $C_i$ , subtract the number of points in  $P$  on the side of  $C_i$  of the blue edge and add the number of points in  $P$  on the side  $C_j$  of the blue edge (see the Figure 3.2). The simplicial depth inside  $C_i$  includes

simplicies (triangles) bounded by the blue line segment and points in  $P$  on the side of  $C_i$ . When moving across the blue edge to  $C_j$ , we exit (subtract) one set of triangles and enter (add) a new set of triangles based on the blue line segment and points of  $P$  on  $C_j$  side of the blue edge. If depth on simplex boundaries is required, then the depth on the blue edge is calculated by adding depth in  $C_i$  to the number of points in the side of  $C_j$ ; no query point lies on a simplex boundary when  $P \cup Q$  is in general position. All cells outside the convex hull of  $P$  have depth zero; we

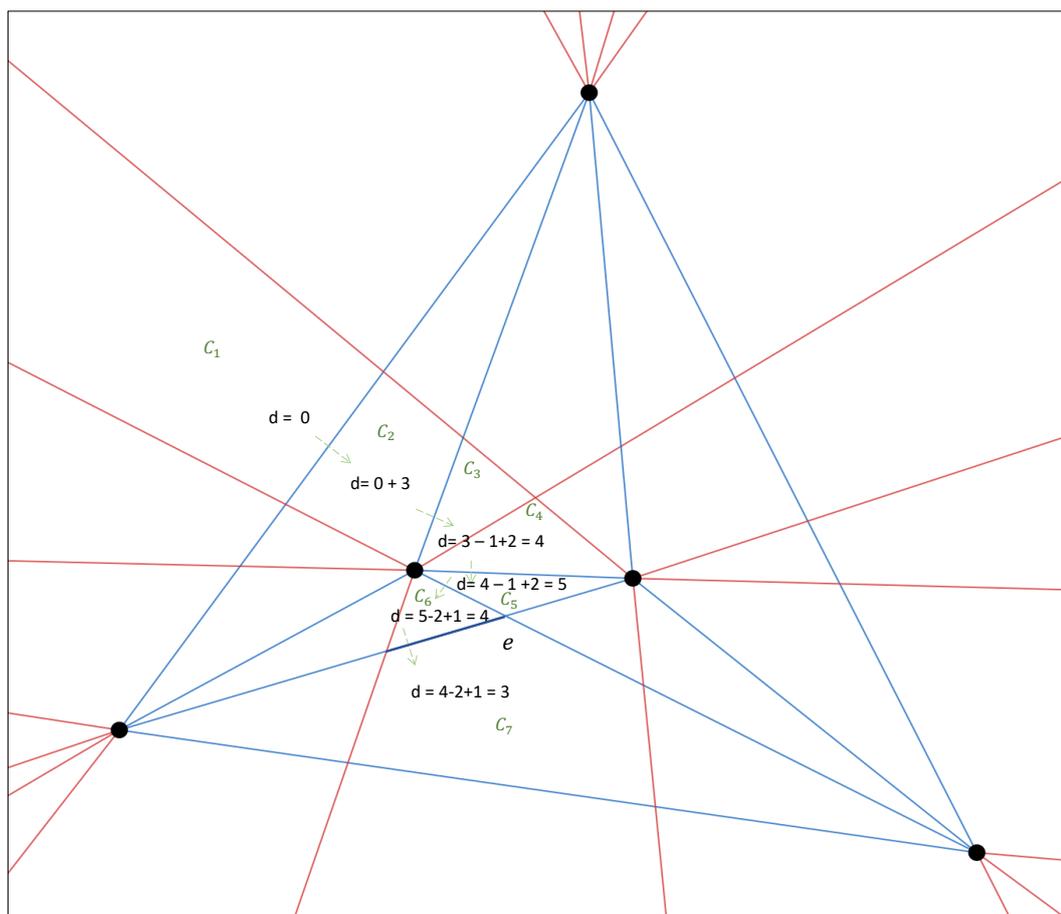


Figure 3.2: The arrangement  $L$  as described with  $n = 5$  points. Traversing  $L$  for computing the depth of each cell can be started from outside the convex hull of  $P$  with zero depth. The depth of a neighboring cell is computed by subtracting the number of points to the left of the shared edge and adding the number of points to the right of the shared edge. For example, to find simplicial depth of  $C_7$  use the simplicial depth of  $C_6$  which share common blue edge  $e$  and has know simplicial depth  $d = 4$ . Now, simplicial depth of  $C_7$  is  $d = 4 - 2 + 1$  ( $2 =$  the number of points in  $P$  on the side of  $C_6$  of  $e$ ,  $1 =$  the number of points in  $P$  on the side  $C_7$  of  $e$ ). One possible path to compute simplicial depths of cells is shown in the figure, we can start for  $C_1$ , which has simplicial depth  $d = 0$  continue above step on  $C_1, C_2, C_3, C_4, C_5, C_6, C_7, \dots$  path.

can initiate our algorithm at any of these cells. The algorithm proceeds to compute the depths of all cells by traversing the planar graph determined by  $L$  starting from an extreme cell (with depth zero) using the above technique. The depth of each individual cell is computed in  $O(1)$  time. Therefore, the traversal takes time and space proportional to the number of cells:  $\Theta(n^4)$ .

Finally, to obtain the depth of each point in  $Q$ , we apply a point location algorithm to locate each point of  $Q$  in its corresponding cell in the arrangement determined by  $L$ . Kirkpatrick's point location algorithm can be implemented in  $t$ -vertex planar subdivision with  $O(t)$  preprocessing time,  $O(t)$  space, and  $O(\log t)$  query time [25]. In our case,  $t \in \Theta(n^4)$ , corresponding to  $\Theta(n^4)$  cells in the planar subdivision determined by  $L$ . Therefore, Kirkpatrick's point location algorithm can be used to find the locations of each point in  $Q$  in  $O(n^4)$  preprocessing time,  $O(n^4)$  space and  $O(k \log n)$  query time. The simplicial depths of all points in  $Q$  can be computed in  $O(n^4 + k \log n)$  time and  $O(n^4)$  space. This gives the following lemma:

**Lemma 3.3.** *Given a set  $P$  of  $n$  points and a set  $Q$  of  $k$  query points in general position in  $\mathbb{R}^2$ , Algorithm S.III computes  $SD_P(q)$  for every  $q \in Q$  relative to  $P$  in  $O(n^4 + k \log n)$  time and  $O(n^4 + k)$  space.*

The algorithms S.I, S.II and S.III discussed above compute the simplicial depths of  $k$  points relative to a set of  $n$  points for all values of  $n$  and  $k$ . Algorithm S.I is asymptotically fastest when  $k \in O(\frac{n}{\log n})$ , Algorithm S.II is asymptotically fastest when  $k \in \Omega(\frac{n}{\log n})$  and  $k \in O(n^3)$ , and Algorithm S.III is asymptotically fastest when  $k \in \Omega(n^3)$ . This gives the following theorem:

**Theorem 1 (Simplicial depth of Batched Query Points).** *Given a set  $P$  of  $n$  points and a set  $Q$  of  $k$  query points in general position in  $\mathbb{R}^2$ ,  $SD_P(q)$  for every  $q \in Q$  relative to  $P$  can be computed in  $O(\min\{kn \log n, n^2 + nk, n^4 + k \log n\})$  time.*

## 3.2 Algorithms for Computing Tukey Depths of Batched Query Points

In this section, we describe two methods for computing the Tukey depths of multiple query points which are based on previous work related to Tukey depth [38] and Tukey

depth contours [32].

### 3.2.1 Algorithm T.I

In  $\mathbb{R}^2$ , the Tukey depth of a point  $q$  relative to a set  $P$  of  $n$  points can be computed in  $O(n \log n)$  time [38]. Similar to Algorithm S.I in Section 3.1, a straightforward method for computing the Tukey depths of  $k$  query point is to apply a Tukey depth algorithm iteratively for each point of  $Q$ . This process take  $O(kn \log n)$  time and  $O(n)$  space to store the sorted order of  $P$  around each point of  $Q$ . This gives the following lemma:

**Lemma 3.4.** *Given a set  $P$  of  $n$  points and a set  $Q$  of  $k$  query points in general position in  $\mathbb{R}^2$ , Algorithm T.I computes  $TD_P(q)$  for every  $q \in Q$  relative to  $P$  in  $O(kn \log n)$  time and  $O(n + k)$  space.*

### 3.2.2 Algorithm T.II

Algorithm T.I is efficient when  $k$  is small relative to  $n$ , but more efficient approaches are possible for larger values of  $k$ . Algorithm T.II begins by computing the Tukey depth contours of  $P$  using the algorithm of Miller et al. in  $O(n^2)$  time and space [32]. Miller et al. showed how to build a point location data structure on the contours in  $O(n^2)$  time to support  $O(\log n)$ -time depth queries. Therefore, the Tukey depths of  $k$  points can be calculated in  $O(n^2 + k \log n)$  time and  $O(n^2)$  space. This gives the following lemma:

**Lemma 3.5.** *Given a set  $P$  of  $n$  points and a set  $Q$  of  $k$  query points in general position in  $\mathbb{R}^2$ , Algorithm T.II computes  $TD_P(q)$  for every  $q \in Q$  relative to  $P$  in  $O(n^2 + k \log n)$  time and  $O(n^2 + k)$  space.*

Algorithms T.I and T.II compute the Tukey depths of  $k$  points relative to a set of  $n$  points for all values of  $n$  and  $k$ . Algorithm T.I is asymptotically fastest when  $k \in O(\frac{n}{\log n})$ , and Algorithms T.II is asymptotically faster when  $k \in \Omega(\frac{n}{\log n})$ . This gives the following Theorem:

**Theorem 2 (Tukey depth of Batched Query Points).** *Given a set  $P$  of  $n$  points and a set  $Q$  of  $k$  query points in general position in  $\mathbb{R}^2$ , computes  $TD_P(q)$  for every  $q \in Q$  relative to  $P$  can be computed in  $O(\min\{kn \log n, n^2 + k \log n\})$  time.*

### 3.3 Application of Batched Depth Queries

Data depth is an important concept in nonparametric data analysis, especially when analyzing multivariate data as it can provide a center-outward ordering of a data set. Here the depths of points in  $P$  are used to rank the points in  $P$ . An important application of batched depth queries is that we can generate a center-outward ordering of points of a set  $Q$  relative to  $P$ . Depth-based ranking leads to the construction of nonparametric tests and of several tools for the analysis of multivariate data sets, such as bagplots [39] and DD plots [28]. One possible application of ranks of points in  $Q$  relative to  $P$  (and perhaps of points in  $P$  relative to  $Q$  as well) may be applying a multivariate rank test to compare  $P$  and  $Q$ , which will need further studies. We emphasize that Li and Liu [28] have introduced several methods to compare locations and scales of multivariate distributions based on data depth. For instance, a DD plot is a two-dimensional graphical tool to compare two multivariate distributions /samples based on data depth [28]. For two random samples  $\{X_1, \dots, X_m\} = \mathbf{X}$  and  $\{Y_1, \dots, Y_n\} = \mathbf{Y}$ , respectively from  $F$  and  $G$  in  $\mathbb{R}^d$ , the DD-plot is the plot of  $DD(F, G)$ , where

$$DD(F, G) = \{D_F(x), D_G(x), x \in \{\mathbf{X} \cup \mathbf{Y}\}\},$$

$D_F(x)$  and  $D_G(x)$  are the simplicial depths of  $x$  with respect to the distributions  $F$  and  $G$ , respectively. If  $F$  and  $G$  are unknown, the DD plot is defined using

$$DD(F_m, G_n) = \{D_{F_m}(x), D_{G_n}(x), x \in \{\mathbf{X} \cup \mathbf{Y}\}\}, \quad (3.1)$$

with  $F_m$  and  $G_n$  denoting the empirical distributions constructed from the  $\mathbf{X}$  and  $\mathbf{Y}$  samples, respectively. Equation 3.1 requires computing the depth of points in the samples  $\mathbf{X}$  and  $\mathbf{Y}$  against both populations independently. An important application of batched simplicial depth queries and of the algorithms we presented in previous sections, is to efficient calculation of depth values for DD plots.

Another application is to build a bagplot, which is a bivariate generalization of the univariate box-plot and was introduced by Rousseeuw et al. [39] based on the Tukey depth. The bagplot consists of the deepest point, the bag (central area)

consists of the 50% most central points, the fence which is found by magnifying the bag by a factor 3, and the loop which is the area between the bag and fence, outliers are recognized as the points outside the fence. One possible application of batched depth computation is in the construction of bagplots, from which we can identify points of  $Q$  with interesting characteristics with respect to  $P$ , such as the deepest query point in  $Q$  or points of  $Q$  in the tails of  $P$ .

# Chapter 4

## Defining the Depth of a Query Set

In the previous chapter, we introduced algorithms to compute batched depth queries of  $k$  query points relative to a set  $P$ . In this chapter, we consider answering the following question: how can we measure the centrality of set  $Q$  relative to set  $P$  based on simplicial depth and Tukey depth? To answer this question, we present new definitions for the simplicial depth and Tukey depth of a set of points in  $\mathbb{R}^d$ . Then, we present a probabilistic interpretation for the new definition of simplicial depth of a query set of points. Next, we present a comparison between the cardinal simplicial depth [6] and the new definition of simplicial depth of a query set of points. In Section 4.3, we introduce a new definition for the Tukey depth of a query set of points. Then, we highlight the differences between other definitions of Tukey depth of sets discussed in Section 2.3.2 and our new definition. Finally, we end this chapter by discussing possible applications of the depth of a set of points introduced here.

### 4.1 Defining the Simplicial Depth of a Query Set

We first define a new notion of simplicial depth of the query set  $Q$  with respect to the input set  $P$  to quantify the centrality of the set  $Q$  relative to the set  $P$ . This depth measure corresponds to the sum of the number of points of  $Q$  contained in each simplex determined by points in  $P$ , normalized by the number of points in  $Q$ :

**Definition 4.1 (Simplicial Depth of a Query Set).** *Given a set  $P$  of  $n$  points and a set  $Q$  of  $k$  points in  $\mathbb{R}^d$ , the simplicial depth of  $Q$  relative to  $P$  is*

$$SD_P^*(Q) = \frac{1}{|Q|} \sum_{S \in \mathcal{S}} |Q \cap S|,$$

where  $\mathcal{S}$  denotes the set of  $\binom{n}{d+1}$  closed simplicies, each of which is the convex hull of  $d + 1$  points from  $P$ .

As a first simple property, we note that the above depth  $SD_P^*(Q)$  can be expressed as the average of the simplicial depths of points in  $Q$ . Recall that  $SD_P(q)$  measures the centrality of a query point  $q$  with respect to  $P$ . Given this, the depth of the set  $Q$  as defined above has a simple interpretation.

**Lemma 4.1.**

$$SD_P^*(Q) = \frac{1}{|Q|} \sum_{q \in Q} SD_P(q)$$

*Proof.* First, from Definition 1.2, we recall that,

$$SD_P(q) = \sum_{S \in \mathcal{S}} I(q \in S)$$

Then, we have that,

$$\begin{aligned} SD_P^*(Q) &= \frac{1}{|Q|} \sum_{S \in \mathcal{S}} |Q \cap S| \\ &= \frac{1}{|Q|} \sum_{S \in \mathcal{S}} \sum_{q \in Q} I(q \in S) \\ &= \frac{1}{|Q|} \sum_{q \in Q} \sum_{S \in \mathcal{S}} I(q \in S) \\ &= \frac{1}{|Q|} \sum_{q \in Q} SD_P(q), \end{aligned} \tag{4.1}$$

□

Upon investigating the two sums in Equation (4.1) above,  $SD_P^*(Q)$  can be computed by first obtaining the simplicial depth  $SD_P(q)$  for each individual point  $q \in Q$ , and then taking the average of these depth values. This can be achieved efficiently using any one of the three algorithms introduced in Section 3.1, which gives the following corollary.

**Corollary 4.1.** *Given a set  $P$  of  $n$  points and a set  $Q$  of  $k$  points in general position in  $\mathbb{R}^2$ ,  $SD_P^*(Q)$  can be computed in  $O(\min\{kn \log n, n^2 + nk, n^4 + k \log n\})$  time.*

Section 2.2.2 mentioned that a simplicial median is at an intersection of simplex boundaries. Yet, depending on the arrangement of  $P$ , the maximum simplicial depth may be achieved by more than one point. Therefore the simplicial median may not lie on a unique point. When the set  $Q$  lies on these points, which all reach the maximum simplicial depth  $SD_P(q)$  for a single query point, the simplicial depth  $SD_P^*(Q)$  will achieve its maximum over all sets  $Q$  of size  $k$ .

In Section 2.3, we discussed the cardinal simplicial depth (Definition 2.1) of a set  $Q$  relative to  $P$  introduced by Barba et al. [6]. Cardinal simplicial depth counts the number of non-empty open simplices determined by  $P$  that contain at least one point of  $Q$ . Next, we compare the cardinal simplicial depth of  $Q$  relative to  $P$  with the simplicial depth  $SD^*$  of  $Q$  relative to  $P$  (Definition 4.1) via the following observations.

1. The previous study by Barba et al. [6] sought to find higher-dimensional analogs to quantiles. In other words, their objective was to find a minimal point set  $Q$ , which represents or summarizes  $P$ ; for that, they defined the cardinal simplicial depth of a set  $Q$  of points relative to the set  $P$ . In contrast, we aim to construct a measure of the depth of an arbitrary given query set  $Q$  relative to another given point set  $P$  that quantifies how globally central  $Q$  is with respect to  $P$ , or how  $Q$  is nested within  $P$ .
2. Cardinal simplicial depth results in similar depth values for significantly different point sets. Specifically, a small subset of points in the set  $Q$  can determine the depth of  $Q$  relative to  $P$ . As a result, equal depth values can be obtained from very different configurations of a set  $Q$  for a given set  $P$ . See Figure 4.1 for an example.
3. Only the points of  $Q$  that lie inside the convex hull of  $P$  count towards the value of cardinal simplicial depth. Whereas Definition 4.1 considers the average depth of points in  $Q$ , points outside the convex hull of  $P$  contribute evenly. For example, both query point sets in Figure 4.1 have the same cardinal simplicial depth but, according to Definition 4.1, the red query set has low depth

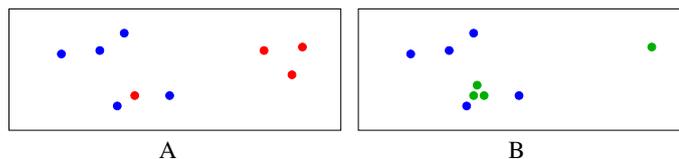


Figure 4.1: The green and red sets have the same cardinal simplicial depth relative to the blue set. The green set's simplicial depth relative to the blue set is triple that of the red set's simplicial depth relative to the blue set by new Definition 4.1. These same examples demonstrate an analogous property for Tukey depth: the green set and the red set have the same generalized Tukey depth (by Definition 2.2) relative to the blue set, but their Tukey depths differ by new Definition 4.2.

compared to the green query set. Another example is shown in Figure 4.2: adding two more red points in  $\triangle ABD$  and  $\triangle BCE$  increases the cardinal simplicial depth of red points up to 10 (the maximum achievable depth in this case), but when more query points are added inside or outside the convex hull of  $P$ , this does not change the cardinal simplicial depth of the red point set. In contrast, every time a new blue point is added the new definition of simplicial depth will lead to a new value of the depth of the blue set unless all points are added to the center region with the previous points.

4. Definition 4.1 is a normalized sum of the number of points contained in each simplex. It also suggests a family of measures that can be used to define the simplicial depth of a set  $Q$  with respect to a set  $P$  by substituting the average with another summary statistic of the distribution of the depths of points in  $Q$ . For instance, we could summarize the distribution of depths using a median, a minimum, a maximum, or a measure of spread, such as variance, range, skewness, or quantiles of this distribution. These different summaries of the constructed depth distributions over the points of  $Q$  can all be computed in the same time and space complexities as in Corollary 4.1.
5. The location of  $Q$  in  $P$ , which maximize the cardinal simplicial depth and new simplicial depth of the set, are different. According to Definition 4.1, the maximum simplicial depth for  $Q$  is generally obtained when  $Q$  consists of points located at a simplicial median of  $P$  since a simplicial median is a deepest point in  $P$ . If we consider both  $P$  and  $Q$  in general position, then the maximum  $SD_P^*(Q)$  can be achieved by placing  $Q$  in a deepest simplicial depth cell (for an arrangement of  $P$ , an open cell region bounded by simplices

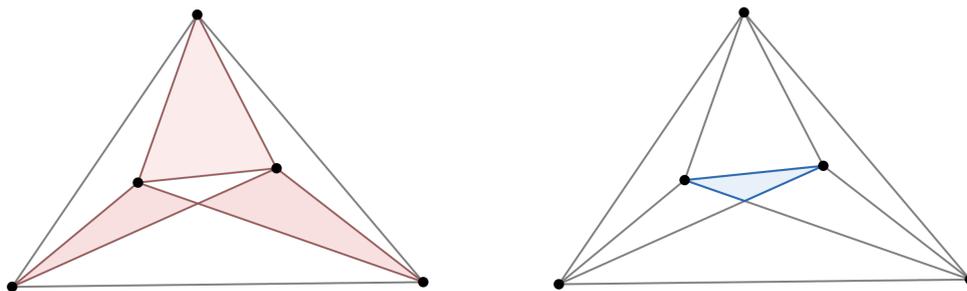


Figure 4.2: For a set  $P$  of 5 points, and  $P \cup Q$  in general position, placing a query point in each of the red cells gives the maximum achievable cardinal simplicial depth (with  $CSD_P(Q) = 8$  out of 10 possible triangles) for a set  $Q$  of 3 points, as in the figure on the left side. Placing three points anywhere in the blue cell gives the maximum simplicial depth according to Definition 4.1 (each point having a maximum of  $SD_P(q) = 5$ ), as in the figure on the right side.

boundaries with the highest simplicial depth is equal for both closed simplicial depth and open simplicial depth). In contrast, to maximize cardinal simplicial depth, points in  $Q$  should be spread within the convex hull of  $P$  in a way as to cover as many simplices as possible formed from the points of  $P$ . See Figure 4.2.

6. In  $\mathbb{R}^2$ , consider a special arrangement in which  $P$  consists of vertices of a regular  $n$ -gon, where  $n$  is even and  $Q$  has two points. Here, we consider both definitions in terms of closed simplices for a fair comparison of cardinal simplicial depth with Definition 4.1. The arrangement of the points of  $Q$  that gives the maximum cardinal closed simplicial depth for two query points, among any other arrangements of those two points, is obtained when the two query points are close to the center of the polygon, and on lines connecting the center and two opposite points in  $P$ . See Figure 4.3. Placing those two points at the center of the  $n$ -gon gives the maximum simplicial depth for the set  $Q$  according to Definition 4.1.

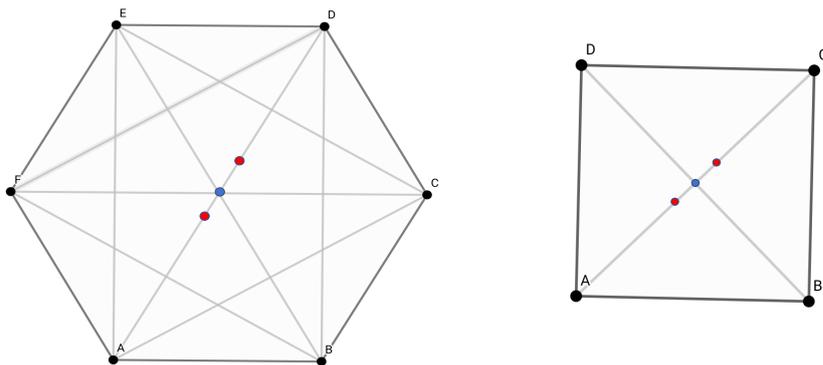


Figure 4.3: The set  $P$  of 6 and 4 points are in the vertices hexagon and square, respectively. The sets made of two points that attain maximum cardinal closed simplicial depth are indicated in red (the two points should be placed close to the center and on the line connecting the center and two opposite vertices). In contrast, two overlapping points at the center of the regular  $n$ -gon gives the maximum simplicial depth according to Definition 4.1.

## 4.2 Probabilistic Interpretation for the Simplicial Depth of a Query Set

Data depth and depth-based applications are discussed in both Computer Science and Statistics. One common method is to define data depth as a function of the coordinates of points in  $P$  and  $Q$  that returns a single value to quantify the centrality of  $Q$  relative to  $P$ . Techniques in computational geometry, such as the algorithms discussed in previous sections, can be used to calculate exact values for data depth. In statistics, data depth is also defined as a probability of an event with distributional characteristics. For example, later in this section, we discuss simplicial depth as an estimator of a probability. Techniques from Computational Statistics, such as Monte Carlo approximation can be used to efficiently calculate an approximate depth value, especially when the dimension of the data set is high, as the exact calculation of depth is computationally expensive in high dimensions. For example, several studies provided approximation algorithms for simplicial depth [12, 1] and approximation algorithms for Tukey depth of a query point in  $\mathbb{R}^d$  were introduced in [40, 15]. In this section, we provide a statistical view for the simplicial depth of a query set to view the problem from a different perspective.

Firstly, we note that Lemma 4.1 implies that  $SD_P^*(Q)$  also has a natural probabilistic interpretation. Consider a point  $q$  selected uniformly at random from  $Q$ .

The expected value of the simplicial depth of  $q$  relative to  $P$  is then  $SD_P^*(Q)$ .

Next, we consider another generalization of simplicial depth to sets, which we show is equivalent to Definition 4.1. For this, we begin by introducing the *normalized simplicial depth* ( $NSD$ ) of a query point  $q$  relative to  $P$  as

$$NSD_P(q) = \frac{1}{|\mathcal{S}|} \sum_{S \in \mathcal{S}} I(q \in S) = \frac{SD_P(q)}{|\mathcal{S}|}, \quad (4.2)$$

with the indicator function  $I$  and the set  $\mathcal{S}$  of simplices defined as in Definition 1.2. Note that this normalized simplicial depth can also be interpreted as the probability that the query point  $q$  lies in a simplex constructed by randomly selecting  $d + 1$  different points from  $P$  or, equivalently, as

$$NSD_P(q) = \mathbb{P}(q \in S), \quad (4.3)$$

where  $S$  is selected uniformly at random from  $\mathcal{S}$ . As seen in (4.2), this also corresponds to the proportion of simplices from  $\mathcal{S}$  that contain  $q$ . Liu [29] argued that (4.3) is an estimator of the probability that the query point  $q$  lies in a simplex formed from  $d + 1$  independent random points selected from a common distribution  $F$  in  $\mathbb{R}^d$ . Finally, we note that  $NSD_P$  and  $SD_P$  are equivalent depth measures as they differ by a normalizing factor and would, for instance, lead to the same conclusion when ranking points with respect to their depths.

Now, consider generalizing the above idea by randomly selecting a simplex from  $\mathcal{S}$ , and focusing on the expected number of points of  $Q$  that lie in that simplex. This depth measure, which we denote  $ERS_P(Q)$  (Expected number of points of  $Q$  in a Random Simplex from  $P$ ) is then,

$$ERS_P(Q) = \mathbb{E}[Y_Q(S)], \quad (4.4)$$

where  $S$  is again randomly selected from  $\mathcal{S}$ , and where the random variable  $Y_Q(S)$  denotes the number of points of  $Q$  that lie inside  $S$ . This is a reasonable measure of the depth of  $Q$  with respect to  $P$ , has an elegant and intuitive interpretation, and reduces to (4.3) when  $Q$  contains a single point. Indeed, when  $Q$  contains a single point,  $Y_Q(S) = I(q \in S)$  and  $\mathbb{E}[Y_Q(S)] = \mathbb{P}(q \in S)$ , the normalized simplicial depth

of  $q$ .

We now proceed to justify that  $ERS_P$  is equivalent to  $SD_P^*$ . The number of points of  $Q$  that lie inside a simplex  $S$  constructed from points of  $P$  can be expressed as

$$Y_Q(S) = \sum_{q \in Q} I(q \in S), \quad (4.5)$$

and takes values in  $\{0, 1, \dots, |Q|\}$ . Also, the proportion of simplices in  $\mathcal{S}$  that have exactly  $y$  points of  $Q$  in them is

$$P_{\mathcal{S}}(y) = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} I[Y_Q(s) = y],$$

for  $y = 0, 1, \dots, |Q|$ . The above also corresponds to the probability that a randomly selected simplex constructed from points of  $P$  contains exactly  $y$  points of  $Q$ . To avoid confusion, in the above and in what follows, we use  $S$  to denote a randomly selected simplex, and we use  $s$  otherwise. In this context, the expectation of  $Y_Q(S)$  corresponds to the mean of the above probability distribution, that is

$$\begin{aligned} \mathbb{E}[Y_Q(S)] &= \sum_{y=0}^{|Q|} y P_{\mathcal{S}}(y) \\ &= \sum_{y=0}^{|Q|} \frac{y}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} I[Y_Q(s) = y] \\ &= \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \sum_{y=0}^{|Q|} y I[Y_Q(s) = y] \\ &= \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} Y_Q(s). \end{aligned}$$

Using (4.5), we can further simplify the above to get

$$\begin{aligned} \mathbb{E}[Y_Q(S)] &= \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \sum_{q \in Q} I(q \in s) \\ &= \frac{1}{|\mathcal{S}|} \sum_{q \in Q} \sum_{s \in \mathcal{S}} I(q \in s) \\ &= \frac{1}{|\mathcal{S}|} \sum_{q \in Q} SD_P(q) \\ &= \frac{|Q|}{|\mathcal{S}|} SD_P^*(Q). \end{aligned} \quad (4.6)$$

Interestingly, (4.4) and (4.6) together imply that

$$SD_P^*(Q) = \frac{|\mathcal{S}|}{|Q|} ERS_P(Q).$$

From this, the simplicial depth of  $Q$ , as it is defined in Definition 4.1, is equivalent to  $ERS_P(Q)$ . As a result,  $SD_P^*$  is equivalent to the expected number of points in  $Q$  that lie in a simplex constructed from points selected at random from  $P$ .

We note that  $CSD_P$  was defined in (2.1) as the number of simplices constructed from points of  $P$  that contain at least one point of  $Q$ . As such, it can be expressed as

$$CSD_P(Q) = \sum_{s \in \mathcal{S}} I[Y_Q(s) > 0].$$

Derivations similar to those provided above allow one to see that

$$\begin{aligned} CSD_P(Q) &= \sum_{s \in \mathcal{S}} \sum_{y=1}^{|Q|} I[Y_Q(s) = y] \\ &= \sum_{s \in \mathcal{S}} \sum_{y=0}^{|Q|} I[y > 0] I[Y_Q(s) = y] \\ &= |\mathcal{S}| \sum_{y=0}^{|Q|} I[y > 0] P_{\mathcal{S}}(y) \\ &= |\mathcal{S}| \mathbb{E}[I(Y_Q(S) > 0)], \end{aligned}$$

or, that  $CSD_P(Q)$  is proportional to  $\mathbb{E}[I(Y_Q(S) > 0)] = \mathbb{P}(Y_Q(S) > 0)$ , the probability that a random simplex contains at least one point of  $Q$ . Obviously, in the case where  $Q$  contains a single point, this further reduces to  $\mathbb{P}(Y_Q(S) > 0) = \mathbb{P}(q \in S)$  and justifies that  $CSD_P(Q)$  is a direct generalization of simplicial depth that applies to sets, but differs from  $ERS_P(Q)$  defined in (4.4).

### 4.3 Tukey Depth of a Query Set

Motivated by our new definition for simplicial depth of a query set  $Q$  relative to a set  $P$  introduced in Section 4.1 and by Lemma 4.1, we introduce a new notion of Tukey depth of a query set  $Q$  relative to a set  $P$ :

**Definition 4.2 (Tukey depth of a Query Set).** *Given a set  $P$  of  $n$  points and a set  $Q$  of  $k$  points in  $\mathbb{R}^d$ , the Tukey depth of  $Q$  relative to  $P$  is*

$$TD_P^*(Q) = \frac{1}{|Q|} \sum_{q \in Q} TD_P(q). \quad (4.7)$$

As with the simplicial depth of a set  $Q$  relative to  $P$ , (4.7) corresponds to the average depth of points in  $Q$  relative to  $P$ , and carries the same probabilistic interpretation as for simplicial depth; that is, it corresponds to the expected depth of a randomly selected point of  $Q$  (all points of  $Q$  being equally likely to be selected).

In order to obtain  $TD_P^*(Q)$ , one can calculate the Tukey depth of each of the  $k$  points of  $Q$  with respect to  $P$  using the methods described in Section 3.2, and take the average of those depth values. Therefore, we have the following corollary.

**Corollary 4.2.** *Given a set  $P$  of  $n$  points and a set  $Q$  of  $k$  points in general position in  $\mathbb{R}^2$ ,  $TD_P^*(Q)$  can be computed in  $O(\min\{kn \log n, n^2 + k \log n\})$  time.*

The Tukey depth of a single query point is a minimization problem, because it finds the minimum number of points in  $P$  in any halfspace containing the query point. Defining the Tukey depth of a query point as a probability of an event as with Liu's definition for simplicial depth of a single query point has not been discussed. Therefore, we do not obtain an interpretation for  $TD_P^*(Q)$  analogous to the relationship between  $SD_P^*(Q)$  and  $ERS_P(Q)$ . But, as mentioned above, Definition 4.2 can also be interpreted as the mean or the expected depth of a randomly selected point from  $Q$ .

We mentioned in Section 2.3.2 a generalized Tukey depth of a set  $Q$  relative to a set  $P$ ,  $GTD_P(Q)$ , introduced by Pilz and Schneider [35]. The idea of their study was to extend and find quantiles in a higher dimensional data set. Our goal here is different: we focus on quantifying the centrality of a query point set relative to another set. In one-dimensional space, for every ray containing one of 1/3-quantile and 2/3-quantile contain one-third of the data set, a ray containing both quantiles contains two-thirds of the data set [35]. Generalizing this idea, Pilz and Schneider defined  $GTD_P(Q)$  as the minimum ratio between the number of points in  $P$  and the number of points in  $Q$  contained in any halfspace which contains at least one point of  $Q$ . Similar to the cardinal simplicial depth,  $GTD_P(Q)$  can result in similar depth

values for significantly different point sets. Specifically, a small subset of points in  $Q$  can determine the depth of  $Q$  relative to  $P$ . See Figure 4.1. Pilz and Schneider did not provide an algorithm to compute  $GTD_P(Q)$ , but based on Definition 2.2, a straightforward iterative approach for computing  $GTD_P(Q)$  in  $\mathbb{R}^2$  would require  $O(n^3 + k^3)$  time to compute  $O(n^2 + k^2)$  halfspaces and to count the number of points from  $P$  and  $Q$  in each halfspace. This time can likely be reduced to  $O(n^2 + k^2)$  by constructing the arrangement of lines determined by pairs of points in  $Q \cup P$  and traversing the arrangement to examine all possible subsets of  $Q \cup P$  contained in a half-plane; traversing from one cell in the arrangement to a neighbouring cell corresponds to adding or removing  $O(1)$  points from  $Q \cup P$ .

Further, in Section 2.3.2, we mentioned Affine Tukey depth ( $ATD_P(Q)$ ) and convex Tukey depth ( $CTD_P(Q)$ ) of a set  $Q$  of query points relative to another set  $P$  of points which were introduced by Bertschinger et al. [8].  $ATD_P(Q)$  was specifically used to define the Tukey depth of a set  $Q$  of  $k + 1$  points,  $k < d$ , and was defined as the minimum number of points of  $P$  in any closed halfspace containing the  $k$ -flat determined by  $Q$ . Because of the restriction on the size of  $|Q|$ , in  $\mathbb{R}^2$ , this definition can only be used to define Tukey depth of sets with  $|Q| = 1$  or  $|Q| = 2$ . Obviously, this definition does not generalize easily to measure the depth of a set with an arbitrary number of query points. On the other hand,  $CTD_P(Q)$  corresponds to the minimum number of points in  $P$  in any closed halfspace containing the convex hull of  $Q$ ; this can be used to define the depth of any  $Q$  without any size restriction on  $Q$ . This definition also has several issues, especially when some points of  $Q$  lie outside the convex hull of  $P$  (when convex hull of  $Q$  intersects the convex hull of  $P$ ). One concern with convex Tukey depth is that the depth of  $Q$  can change significantly with small perturbations of the location of a query point. See Figure 4.4. For another issue, consider the interior points of  $Q$ . According to the definition of  $CTD_P(Q)$ , the number of points and combinatorial structure of these interior points of  $Q$  does not essentially affect the depth of  $Q$ ; in fact, the depth of  $Q$  can only be determined by points on its convex hull.

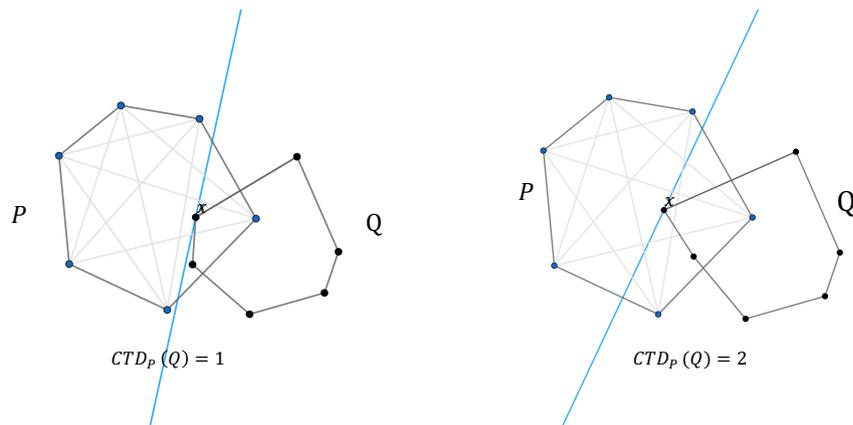


Figure 4.4: The two images give an example for when the convex-hulls of  $P$  and  $Q$  are intersecting. In both examples, the convex Tukey depth of  $Q$  relative to  $P$  is determined by the point  $x$  in  $Q$ . Here, the blue line defines a halfspace which determines the convex Tukey depth of  $Q$ .

## 4.4 Applications of Depth of Sets

In previous sections, we described how to derive a single estimator for the depth of a query point set  $Q$  relative to a set  $P$  in order to evaluate the centrality of the set  $Q$  relative to  $P$ . We foresee such a measure playing an important role in certain contexts. For example, in group sports like cycling and soccer, we may need to analyze the position of one group of athletes relative to other athletes. The depths of different subgroups of athletes can be used access the relative position of that group, or the depth of a team with respect to other athletes can perhaps be used to infer the team's strategy.

Classification is another widely discussed problem in Computer Science and Statistics in which various solutions have been presented. The objective of a typical classification problem is to assign a given object to one of many classes with predetermined characteristics. Techniques based on data depth are popular classification techniques applied in Statistics, especially in nonparametric settings because depth-based techniques do not depend directly on distance measures or any distributional assumption. A “maximum depth-based classifier” is a commonly used method based on data depth. This approach classifies a new point as belonging to the class which gives the maximum depth for the new point. See more detailed explanations in [21]. We identify another possible application of depth of a set in group classification: suppose we want to classify a new set of points  $Q$  which comes from the same distri-

bution (with similar characteristics) as one of the given sets of points  $(P_1, \dots, P_i, \dots)$ , each corresponding to one of the predetermined classes. In order to assign  $Q$  to one of these classes, we can compare the depths of  $Q$  with respect to each of the other sets, and then we can classify  $Q$  belonging to the class associated with the set  $P_i$  that gives the maximum depth for set  $Q$ .

Control charts are useful tools for monitoring and controlling a manufacturing process. In these tools, depth-based nonparametric multivariate analysis techniques have been found to be quite attractive, as they do not require any multivariate distributional assumptions. Liu [30] used simplicial depth to construct control charts for monitoring processes associated to multivariate quality measurements. Relying on depth-based ranks, those control charts essentially reduce multivariate data to univariate data, thus allowing for the construction of multivariate control charts by following the same principles used in univariate charts. We close this chapter by outlining how this is done.

A control chart consists of critical values, a lower control limit (LCL) and an upper control limit (UCL). Between the UCL and LCL is the center line (CL). In a control chart, samples from the manufacturing process are recorded in time order, and their measurements are plotted in the control chart. The region outside the LCL and UCL band is called the out-of-control region; if a sample point is plotted outside the out-of-control region, it indicates that the process is out of control. A process is considered to be in control if the samples meet pre-defined quality requirements or characteristics (associated with the LCL and UCL). The process involves two types of samples. One is the reference sample  $Y$ , produced when the process is in control. The other sample denoted  $X$ , is a new set of observations from the manufacturing process to be tested for quality. Liu [30] introduced  $r$  charts,  $Q$  charts and  $S$  charts for multivariate data corresponding to  $X$  charts,  $\bar{X}$  charts and CUSUM charts in the univariate case.  $r$  charts and  $S$  charts require calculating the depths of points in  $X$  relative to  $Y$ . For example, the  $r$  chart plots the relative rank based on the depth of each sample in  $X$  against the sampling order. Here we can use our batched depth algorithms discussed in Chapter 3 to calculate the depth of each point of  $X$  relative to  $Y$  efficiently. Again, depending on the relative sizes of  $X$  and  $Y$ , we can apply one of the algorithms S.I, S.II, or S.III. As an added benefit, Algorithm S.III constructs

a point location data structure that supports subsequent simplicial depth queries for any query point relative to  $P$  (not restricted to points in  $Q$ ) in  $O(\log n)$  time per query, where  $n = |P|$ . Obviously, this suggests that the concept of monitoring the depth of sets of measurements (as opposed to monitoring the depths of individual measurements) is a natural generalization of current depth-based control charts. Any advances in characterizing the depth of sets would then be directly applicable in the current context. A detailed discussion on the process of identifying a shift in a manufacturing process is presented in [30].

# Chapter 5

## Properties for Depth Measures of Sets of Points

Various properties and characteristics of depth measures are commonly considered. These properties provide a basis for comparing depth measures, allowing one to choose suitable depth measures for a particular application, such as center outward ordering. Further, these properties can help understand the depth characteristics of a whole set of points, the behavior of the depth of a point under some perturbations of  $P$ , or the features of depth related estimators such as the median.

Liu [29] first introduced several desirable properties for data depth measures: affine invariance (P1), maximality at center (P2), monotonicity relative to the deepest point (P3), and vanishing at infinity (P4) (see Sections 5.1 - 5.4). Later, Zuo and Serfling [45] introduced a general definition of a “statistical depth function” based on these four properties and investigated several depth measures according to these properties. This chapter discusses these properties and several other properties commonly discussed with depth measures, which we label P1 - P7.

We examine whether these properties can be adapted or generalized to the case of the depth of a set of query points. We discuss several generalizations of these properties for the depth of sets of points, including cardinal simplicial depth ( $CSD_P(Q)$ ), generalized Tukey depth ( $GTD_P(Q)$ ), convex Tukey depth ( $CTD_P(Q)$ ), and affine Tukey depth ( $ATD_P(Q)$ ). We establish which of these properties are satisfied (or not) by the new notion of simplicial depth for a query set of points ( $SD_P^*(Q)$ ) and the new notion of Tukey depth for a query set of points ( $TD_P^*(Q)$ ) introduced in this

thesis, and by the other generalizations (see the summary in Table 5.2.). Because  $ATD_P(Q)$  (see Definition 2.3) does not generalize to measure the depth of a set of query points with an arbitrary number of query points, we do not examine it further.

We discuss P2, P3 and their generalizations with central symmetric configurations of  $P$  among several possible symmetries. Let  $P$  is a set of points  $\mathbb{R}^d$  and  $c \in \mathbb{R}^d$  be a point. The set  $P' = 2c - P$  is said to be *centrally symmetric to  $P$  with respect to the center  $c$* . In other words,  $P'$  is centrally symmetric to  $P$  with respect to the center  $c$  if and only if for every point  $x \in P$  there is a point  $x' \in P'$  such that  $\frac{1}{2}(x + x') = c$  and for every point  $y' \in P'$  there is a point  $y \in P$  such that  $\frac{1}{2}(y + y') = c$  [9]. Finally, a set  $P$  is said to be a *centrally symmetric set* if  $P$  is centrally symmetric to itself with respect to some center  $c$ .

Finally, we note that when the set  $Q$  has only a single query point, all the depth measures mentioned above reduce to their respective definitions of the depth of a single query point. Therefore, when  $Q$  has only a single query point, these depth measures defined for a query set of points satisfy (or not) P1 - P7 in the same way the depth they seek to generalize does (or not).

## 5.1 P1 - Affine Invariance

This property verifies whether the depth of a point computed by the given depth measure is invariant under affine transformations. If a depth measure is affine invariant, this implies that the depth of a point does not depend on the underlying coordinate system, and in particular, the depth of that point is not affected by the scale of underlying measurements [45].

**Definition 5.1 ( P1 - Affine Invariance [45]).** *For any set  $P$  of  $n$  points in  $\mathbb{R}^d$ , any point  $q$  in  $\mathbb{R}^d$ , and any affine transformation  $f : \mathbb{R}^d \times \mathbb{R}^d$ , a depth measure  $D(\cdot, \cdot)$  is invariant under affine transformation if*

$$D(q, P) = D(f(q), f(P)),$$

where  $f(P) = \{p' \mid p' = f(p) \text{ for some } p \in P\}$ .

Affine transformations include translation, scaling, reflection, rotation, shear

mappings, etc. Liu [29] showed that both simplicial depth and the associated simplicial median are affine invariant. Tukey depth and the Tukey depth median are also affine invariant [45]. We generalize this property for the depth of a set of points as follows.

### P1.1 - Affine Invariance for Depth Measures of Sets

A depth measure  $D(\cdot, \cdot)$  defined for sets is invariant under affine transformation if,

$$D(Q, P) = D(f(Q), f(P)),$$

for any affine transformation  $f$ . In other words, the depth of a point set  $Q$  relative to a set  $P$  should not depend on the underlying coordinate system.

We introduced new definitions  $SD_P^*(Q)$  (Definition 4.1) and  $TD_P^*(Q)$  (Definition 4.2) of a set  $Q$  of points as the average depth of points in  $Q$  relative to a point set  $P$ . These new definitions of depth of a set are linear combinations of the depth of each point in  $Q$ . Since simplicial depth ( $SD_P(q)$ ) and Tukey depth ( $TD_P(q)$ ) of one query point are affine invariant, it immediately follows that  $SD_P^*(Q)$  and  $TD_P^*(Q)$  also satisfy P1.1. Similarly, P1.1 is satisfied for  $CSD_P(Q)$ ,  $GTD_P(Q)$ , and  $CTD_P(Q)$  for any set  $Q$  of points.

## 5.2 P2 - Median at the Center of Symmetry

This property reflects that, for a distribution with a uniquely defined center of symmetry, a depth function should attain its maximum value at this center.

**Definition 5.2 (P2 - Median at the Center of Symmetry [45]).** *For any set  $P$  of  $n$  points in  $\mathbb{R}^d$  with a unique point of central symmetry  $c$ , a depth measure  $D(\cdot, \cdot)$  has its median at  $c$  if*

$$D(c; P) = \sup_{x \in \mathbb{R}^d} D(x, P).$$

The Tukey depth of a single query point satisfies property P2 [45]. It is unknown whether P2 is satisfied for simplicial depth with respect to centrally symmetric configurations of  $P$  [18]. Liu [29] showed that simplicial depth satisfies P2 for continuous

angularly symmetric distributions, and Zuo and Serfling [45] discussed P2 for simplicial depth with other symmetric distributions but not for centrally symmetric discrete distributions of  $P$ . We extend this property for depth of a set of points as follows.

### P2.1 - Maximality at the Center of Symmetry for Depth Measures of Sets

For a given set  $P$  having a uniquely defined center of central symmetry, the depth of  $Q$  should attain its maximum value when all points in  $Q$  coincide at this center.

Since it is unknown whether simplicial depth ( $SD_P(q)$ ) satisfied P2 when  $P$  is centrally symmetric, it remains to be determined whether  $SD_P^*(Q)$  satisfies P2.1 for centrally symmetric  $P$ . As discussed in Section 4.1,  $CSD_P(Q)$  will not attain its maximum by placing all points of  $Q$  at the center of symmetry of  $P$  when  $P$  is the vertices of a regular  $n$ -gon. Therefore, cardinal simplicial depth fails to satisfy P2.1.

When  $P$  is centrally symmetric and the points of  $Q$  coincide at the center of symmetry of  $P$ ,  $TD_P^*(Q)$  attains its maximum at the center of symmetry of  $P$  since Tukey depth satisfied P2 for a single query point ( $TD_P(q)$ ). Therefore, our new definition for the Tukey depth of a query set of points ( $TD_P^*(Q)$ ) satisfies P2.1. Finally,  $CTD_P(Q)$  also clearly satisfies property P2.1 but, this property needs to be further studied for  $GTD_P(Q)$ .

## 5.3 P3 - Monotonicity Relative to the Deepest Point

It may seem desirable that, when a point moves away from the deepest point along any fixed ray through that point, the depth values along the ray should decrease monotonically [45].

**Definition 5.3 (P3 - Monotonicity relative to the deepest point[45]).** *For any set  $P$  of  $n$  points in  $\mathbb{R}^d$  having deepest point  $c$ , a depth measure  $D(\cdot, \cdot)$  satisfies monotonicity relative to the deepest point when*

$$D(x; P) \leq D(c + \alpha(x - c); P),$$

for any  $\alpha \in [0, 1]$  and  $x \in \mathbb{R}^d$ .

Tukey depth satisfies monotonicity relative to the deepest point [45]. Liu [29] showed that simplicial depth satisfies this property for continuous angular symmetric distributions, but that it fails to satisfy this property for discrete centrally symmetric distributions [45] and here, also for centrally symmetric sets  $P$ . In what follows, we briefly discuss a generalization of P3 when measuring the depth of a set of points.

### **P3.1 - Monotonicity Relative to the Deepest Point for Depth Measures of Sets**

For a given set  $P$  of  $n$  points in  $\mathbb{R}^d$  having deepest point  $c$ , and for any configuration of  $Q$  within  $P$ , when moving a point  $q \in Q$  along a fixed ray from  $c$ , the depth of  $Q$  should not increase.

As the simplicial depth of a single query point fails to satisfy P3, the new simplicial depth of a set ( $SD_P^*(Q)$ ) also fails to satisfy P3.1.  $CSD_P(Q)$  also fails to satisfy P3.1, as moving one point away from a median may increase the total number of simplices formed by  $P$  containing the query points of  $Q$ .

Since moving one point  $q \in Q$  along a ray from the deepest point cannot increase  $TD_P(q)$  for the point  $q$ , when moving  $q$  the average depth of points of  $Q$  also cannot increase. Therefore, our new definition of Tukey depth for a set ( $TD_P^*(Q)$ ) satisfies P3.1. This property needs to be further studied for  $GTD_P(Q)$  and  $CTD_P(Q)$ .

## **5.4 P4 - Vanishing at Infinity**

This property is satisfied by a depth measure if the depth of a query point  $q$  will attain zero when that point is moved arbitrarily far away from the point set  $P$ .

**Definition 5.4 (P4 - Vanishing at Infinity[45]).** For any set  $P$  of  $n$  points in  $\mathbb{R}^d$  and for any  $q \in \mathbb{R}^d$ , the depth measure  $D(\cdot, \cdot)$  vanishes at infinity when

$$D(q; P) \rightarrow 0 \text{ as } \|q\| \rightarrow \infty,$$

where  $\|q\|$  is any norm of  $q$  on  $\mathbb{R}^d$ .

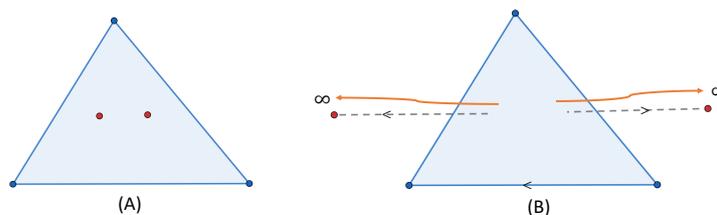


Figure 5.1: An example of sets  $P$  and  $Q$  in  $\mathbb{R}^2$ , respectively, of 3 and 2 points, showing that property P.4.1 is not satisfied for the convex Tukey depth. In (A), points in  $Q$  are interior to the triangle formed by the points in  $P$  and, in (B), the two points in  $Q$  are moving to infinity along fixed rays opposite to each other and parallel to bottom edge of the triangle. In both cases, the  $CTD_P(Q)$  is 1.

Both simplicial depth ( $SD_P(q)$ ) and Tukey depth ( $TD_P(q)$ ) satisfy this property for a single query point [29, 45]. We generalize this property for depth measures of a set of point as follows.

#### P4.1 - Vanishing at Infinity for Depth Measures of Sets

When the points of the query set  $Q$  are moved arbitrarily away from the set  $P$ , the depth of the set  $Q$  relative to  $P$  should approach zero, that is,

$$D(Q; P) \rightarrow 0 \text{ as } \min_{q \in Q} \|q\| \rightarrow \infty.$$

Clearly, moving the set  $Q$  of query points arbitrary away from the other set  $P$  will make the depth of each query point zero, so the average depth of the elements of  $Q$  will be zero. Therefore,  $SD_P^*(Q)$  and  $TD_P^*(Q)$  of a query set  $Q$  relative to a set  $P$  satisfy P4.1. Similarly, P4.1 is satisfied for  $CSD_P(Q)$  and  $GTD_P(Q)$ . But  $CSD_P(Q)$  fails to satisfy P3.1. This is seen from the following counter-example. Let the set  $P$  consist of three points in  $\mathbb{R}^2$  and let the set  $Q$  consist of two points. When  $Q$  is inside the triangle formed by points in  $P$ , the  $CTD_P(Q)$  is 1. Now, move the points of  $Q$  outside the triangle formed by points in  $P$  along two fixed rays opposite to each other and parallel to the bottom edge of the triangle made from points in  $P$  as showed in Figure 5.1. It is then clear that all the halfspaces containing both points in  $Q$  contain at least one point of  $P$  and the convex Tukey depth of  $Q$  remains 1, no matter how far the points of  $Q$  are from the points of  $P$ . Therefore,  $CTD_P(Q)$  fails to satisfy P4.1.

## 5.5 P5 - Consistency Across Dimensions

P5 is satisfied for a depth measure if the depth of a query point is not dependent on the dimension in which the depth was computed. In other words, the depth of a query point  $q$  under the  $k$  dimensional definition should be same as that of the depth under the  $d$  dimensional definition for all  $k < d$  when  $p \cup P$  lies in a  $k$ -flat of  $\mathbb{R}^d$  [18]. Simplicial depth fails to satisfy P5, but Tukey depth does [18]. We generalize P5 for the depth of a set of points as follows.

### P5.1 - Consistency Across Dimensions for Depth Measures of Sets

The depth of a set  $Q$  of query points under dimensions  $k$  and  $d$  ( $k < d$ ) should be equal for all  $k < d$  when  $P \cup Q$  lies in a  $k$ -flat of  $\mathbb{R}^d$ .

We note that  $SD_P^*(Q)$  does not satisfy P5.1 because  $SD_P(q)$  fails to satisfy P5 and  $SD_P^*(Q)$  is the average depth of points in  $Q$ . Similarly,  $CSD_P(Q)$  also does not satisfy the above property.

We briefly turn our attention to Tukey depth and consider  $P$  and  $q$  that are in a  $k$ -flat of  $\mathbb{R}^d$  (for  $k < d$ ). According to the definition of Tukey depth in  $\mathbb{R}^k$ , the halfspace that gives the minimum number of points in  $P$  and containing the query point  $q$  can be extended to a  $d$ -dimensional halfspace. This new halfspace gives the minimum number of points of  $P$  in any closed  $d$ -dimensional halfspace that also contains the query point  $q$ . That is, when  $P$  and  $q$  are in a  $k$ -flat, Tukey depths are equal under the  $k$ - and  $d$ - dimensional definitions. Therefore, when  $P \cup Q$  lies in a  $k$ -flat of  $\mathbb{R}^d$ , the depths of  $TD_P^*(Q)$  are equal whether they are computed according to  $d$ - or  $k$ - dimensional definitions and  $TD_P^*(Q)$  satisfies P5.1. Similarly,  $GTD_P(Q)$  and  $CTD_P(Q)$  also satisfy P5.1.

## 5.6 P6 - Convexity of Depth Contours

This property verifies whether the depth contours of a certain depth measure which bound adjacent regions of different depths are convex [18]. The Tukey depth forms depth contours as a nested collection of convex sets for a set  $P$  and satisfies this property [18, 41] whereas, the simplicial depth ( $SD_P(q)$ ) fails to satisfy this property [45, 18]. P6 is a property associated with the depth characterization of the set  $P$

and describes the characteristics of the depth of a single query point relative to its position in  $P$ . It is not clear how to usefully generalize this property to describe the depth of a set of points.

## 5.7 P7 - Breakdown Point

The breakdown point is an important property that has been widely studied as a measure of the robustness of estimators. Loosely speaking, the breakdown point of a location estimator is the proportion of points that must move arbitrarily far from the rest of the point set in order to make that estimator also move away from the unperturbed points [31]. For example in  $\mathbb{R}^1$ , the median has a breakdown point of  $\lfloor \frac{n-1}{2} \rfloor \frac{1}{n}$  and the mean has a breakdown point of  $1/n$ , for a data set with  $n$  points. This property for various estimators was studied, among many others, by Lopuhaa and Rousseeuw [31] and Donoho and Gasko [17]. In Section 2.2, we explained the use of the breakdown point to describe the robustness of a median. The breakdown point of the Tukey median is between  $\frac{n}{d+1}$  and  $\frac{n}{3}$  [3], and that of the simplicial median is yet unknown [18]. This property itself describes a characteristics relative to the set  $P$ . Therefore, it is not clear how to usefully generalize this property to describe the depth of a set of points.

## 5.8 Summary of Properties

The following two tables summarize the properties we discussed above for depth measure. Table 5.1 summarises the above properties related to simplicial depth and Tukey depth of a single query point. Note that, in the table, P2 and P3 are mentioned relative to a centrally symmetric set  $P$  rather than relative to a continuous centrally symmetric distribution for which the properties do hold. Table 5.2 summarises the above properties for notions of depth for sets of points. Again P2.1 and P3.1 are also mentioned relative to centrally symmetric sets  $P$ .

Zuo and Serfling [45], defined a depth measure satisfying P1 - P4, as a *statistical depth function*. Further, they showed that Tukey depth is a statistical depth function according to their definition and it behaves well overall compared to other depth measures. Additionally, Liu [29] also showed that simplicial depth is a statis-

Table 5.1: Comparing properties of simplicial depth and Tukey depth

	Property	$SD_P(q)$	$TD_P(q)$
P1	affine invariance	✓	✓
P2	median at the centre of symmetry	unknown	✓
P3	monotonicity relative to the deepest point	✗	✓
P4	vanishing at infinity	✓	✓
P5	consistency across dimensions	✗	✓
P6	convexity of depth contours	✗	✓
P7	breakdown point of the median	unknown	1/3

Table 5.2: Comparing properties for depth measure of a set of points

	Property	$SD_P^*(Q)$	$CSD_P(Q)$	$TD_P^*(Q)$	$GTD_P(Q)$	$CTD_P(Q)$
P1.1	affine invariance	✓	✓	✓	✓	✓
P2.1	maximality at the centre of symmetry	unknown	✗	✓	unknown	✓
P3.1	monotonicity relative to the deepest point	✗	✗	✓	unknown	unknown
P4.1	vanishing at infinity	✓	✓	✓	✓	✗
P5.1	consistency across dimensions	✗	✗	✓	✓	✓

tical depth function for continuous angular symmetric distributions. However, for discrete distributions, P2 and P3 are not satisfied by simplicial depth [45]. As we explained above, our new generalization of Tukey depth of a set ( $TD_P^*(Q)$ ) satisfies generalizations of P1 - P4 for depth measures sets. Therefore,  $TD_P^*(Q)$  as introduced in Definition 4.2, is a statistical depth function in the sense of their definition. On the other hand,  $SD_P^*(Q)$  fails to satisfy some properties of P1.1 - P4.1.

# Chapter 6

## Conclusion and Directions for Future Research

This chapter summarizes the main findings of the thesis and suggests possible avenues for future research.

### 6.1 Batched Depth Queries

In this thesis, we studied two interesting questions that arise with depth measures which are relatively new topics. Firstly, Chapter 3 discussed the problem of computing the depths of batched query points. We introduced algorithms for computing the depths of batched queries for two of the most prominent depth measures, simplicial depth and Tukey depth. In  $\mathbb{R}^2$ , for computing the simplicial depths of  $k$  query points relative to a set  $P$  of  $n$  points, we introduced three algorithms: Algorithm S.I which takes  $O(kn \log n)$  time and  $O(n + k)$  space, Algorithm S.II which takes  $O(n^2 + nk)$  time and  $O(n^2 + k)$  space, and Algorithm S.III which takes  $O(n^4 + k \log n)$  time and  $O(n^4 + k)$  space. Similarly, In  $\mathbb{R}^2$ , for computing the Tukey depths of  $k$  query points relative to a set  $P$  of  $n$  points, we introduced two algorithms: Algorithm T.I which takes  $O(kn \log n)$  time and  $O(n + k)$  space and Algorithm T.II which takes  $O(n^2 + k \log n)$  time and  $O(n^2 + k)$  space.

When computing batched depth queries, depending on the relative cardinality of the query point set  $Q$  to the input set  $P$ , one of these algorithms may result in low running time compared to other algorithms. In particular, in computing

batched simplicial depths, Algorithm S.I is suitable when  $k \in O(\frac{n}{\log n})$ , Algorithm S.II performs well when  $k \in \Omega(\frac{n}{\log n})$  and  $k \in O(n^3)$ , and Algorithm S.III is most efficient when  $k \in \Omega(n^3)$ . Similarly, we showed that in computing batched Tukey depths, Algorithm T.I performs better when  $k$  is small relative to  $n$ , i.e., when  $k \in O(\frac{n}{\log n})$ . Otherwise, Algorithm T.II performs in lower running time when  $k \in \Omega(\frac{n}{\log n})$ .

Dynamic algorithms for computing depth are open problems related to depth measures. Our results from the previous sections suggest the following related research problem.

**Open Problem 1:** Can the algorithms designed for batched depth queries be applicable to address dynamic depth problems?

Dynamic depth problems can be viewed in a few different ways.

1. Dynamically maintain the depth of a query point  $q$  / or a set of query points  $Q$ , when new input points are added or removed to the set  $P$ .
2. Dynamically maintain the depths of all points in  $P$  when new input points are added or removed to the set  $P$ .
3. Dynamically compute the depth of a new query point or of a query set with a newly added point.

We suggest that our algorithm for batched depth queries can be useful in designing such dynamic algorithms. Here, we provide one such example, where Algorithm S.II described in Section 3.1.2 can be applied for dynamically maintaining the simplicial depth of a query point when a new point is inserted into  $P$ .

### 6.1.1 Dynamic Simplicial Depth

For this, we maintain a graph  $G$  corresponding to the dual line arrangement of  $P$  ( $L_P$ ) as in Algorithm S.II. Additionally, the dual line  $L_q$  of query point  $q$  is inserted to  $G$ . When a new point  $p' \in \mathbb{R}^2$  is inserted to  $P$ , the dual line  $L_{p'}$  is also inserted to  $G$ , and the doubly connected edge list of  $G$  is updated accordingly. See Figure 6.1. This insertion step takes  $O(n)$  time. Then by traversing  $G$ , the intersections of  $L_{p'}$

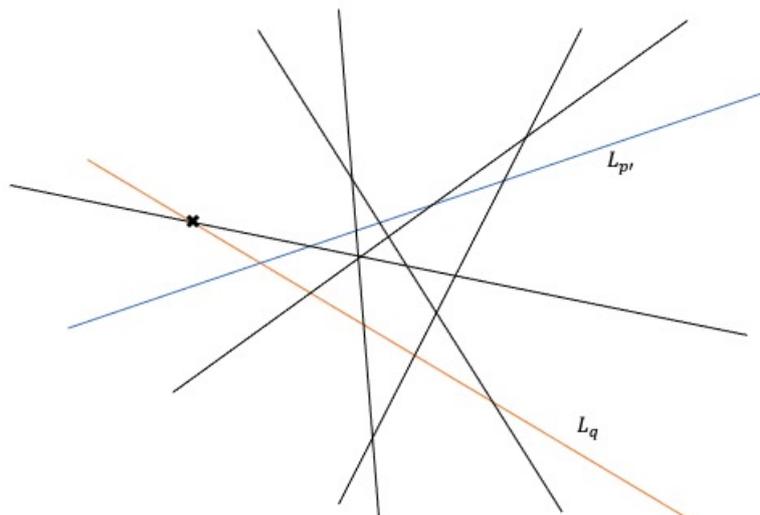


Figure 6.1: Points in the primary plane  $(q, p_1 \dots p_n)$  are transformed to a set of lines  $(L_q, L_1 \dots L_n)$  in the dual plane. Then a new dual line  $L_{p'}$  is added to  $G$ . Finally, line intersection order in  $L_q$  is recalculated including the intersection between  $L_{p'}$  and  $L_q$  [27].

and  $L_q$  and the angular order of the updated  $P$  can be found in  $O(n)$  time. For finding the depth of  $q$ , we can then apply the next step of the simplicial depth algorithm presented by Aloupis et al. [4] (described in Section 2.3.3) in  $O(n)$  time. This whole process requires  $O(n^2)$  space to store  $G$ . When a point  $p' \in P$  is deleted, we can update the simplicial depth of  $q$  by removing  $L_{p'}$  from  $G$  and recalculating the simplicial depth in  $O(n)$  time.

Further, we can use  $G$  to calculate the depth of a new query point  $q'$  as in Algorithm S.II. First  $q'$  is inserted to  $G$ . Next, we can traverse  $G$  to find the sorted intersection order of  $P$  around  $q'$ . Then, using the simplicial depth algorithm presented by Aloupis et al. [4], the simplicial depth of  $q'$  can be computed in  $O(n)$  time.

We discussed batched depth queries with Tukey depth and simplicial depth. There are many other different depth measures such as convex hull peeling depth [7], Oja depth [34], regression depth [37] and so on. We suggest that one possible future work is to study batched depth query problems for other depth measures. And also, we suggest that one can further study the algorithms presented in Chapter 3 to derive algorithms for batched depth queries for other depth measures. For

example, Algorithm S.II in Section 3.1.2 uses dual transformation-based techniques to find the sorted angular order of  $P$  around each  $Q$  in  $O(n^2)$  time and space. This step of Algorithm S.II may be applicable to find the depths of batched queries for other depth measures. Other depth measures such as perihedral depth and eutomic depth [18] also require this type of sorting step in their algorithms. Therefore, we mention the following topic for a future study.

**Open Problem 2:** Extend the idea of batched depth queries for other depth measures.

In Chapter 3, we discussed various applications of batched depth queries. We highlighted a set of nonparametric data analysis applications based on depth-based ranking (center outward ordering) of a set  $Q$  relative to a set  $P$  of points. These depth-based rankings are used for nonparametric tests to compare distributions and scales of multivariate data sets [28], and construct data visualization tools such as bagplots [39] and DD plots [28] for multivariate data sets. One natural direction for future study is to examine applications of batched depth queries in the construction of such statistical tools.

**Open Problem 3:** Study applications of batched depth queries, with a focus on the efficient construction of statistical tools.

## 6.2 Depth of a Query Set of Points

Secondly, in this thesis, we discussed depth of a query set of points relative to another set of points. In Chapter 2 we mentioned that several studies have introduced different definitions for depth of sets, including cardinal simplicial depth [6], generalized Tukey depth [35], convex Tukey depth [8] and Affine Tukey depth [8]. In Chapter 4, we introduced new definitions for the simplicial depth of a set  $Q$  of  $k$  points, relative to a set  $P$  of  $n$  points in  $\mathbb{R}^d$  as sum of the number of points of  $Q$  contained in each simplex determined by points in  $P$ , normalized by the number of points in  $Q$ . Then, we defined Tukey depth of a set  $Q$  of  $k$  points, relative to a set  $P$  of  $n$  points in  $\mathbb{R}^d$ , as the average depth of points in  $Q$ . We discussed how these

depth measures can be computed in  $\mathbb{R}^2$  by applying the algorithms above, leading to computing times of  $O(\min\{kn \log n, n^2 + nk, n^4 + k \log n\})$  for simplicial depth, and  $O(\min\{kn \log n, n^2 + k \log n\})$  time for Tukey depth. We also presented a comparison with new definitions for a set of points and the previously defined depth of a set of points.

Finally, we discussed several applications of the depth of sets based on the new definitions  $SD_P^*(Q)$  and  $TD_P^*(Q)$ . Applications where evaluating the centrality of a set relative to another set is useful such as classifications can make use of these new definitions. We suggest that the concept of depth of a set will be a useful tool in application, specifically in multivariate statistical analysis.

**Open Problem 4:** Explore applications of depth of a set of query points with a focus on multivariate statistical analysis.

This thesis suggests various possible generalizations of simplicial depth and Tukey depth for measuring the depth of a query set  $Q$ . As the computation of these depth measures involves computing the depth of each point in  $Q$ , we could instead define a depth measure as a function of a different summary of the distributions of the simplicial depths and Tukey depths of individual points of  $Q$  relative to  $P$ . For instance, we could summarize the distribution of depths using a median, a minimum, a maximum, or a measure of spread, such as variance, range, skewness, or quantiles of this distribution. These different summaries of the constructed depth distributions over the points of  $Q$  can all be computed in the same time and space complexities as in Corollaries 4.1 and 4.2. From a statistical perspective, using a variety of summary measures could allow for a more in-depth comparison of two sets  $Q_1$  and  $Q_2$  relative to a set  $P$  and could lead to interesting applications.

**Open Problem 5:** Study the use of different summary statistics and of batched depth queries to define new depth measures for sets of points and to construct new statistical tools.

### 6.3 Properties Defined for Depth Measures of Sets of Points

In Chapter 4, we presented a different statistical interpretation for the simplicial depth of a set of points. Specifically, we examined the relationship between the simplicial depths of individual points in  $Q$  relative to  $P$  and the expected number of query points in  $Q$  that lie in a random simplex constructed from points in  $P$ , defined as  $ERS_P(Q)$ . This idea was inspired by the  $\psi$ -histogram [18], also known as  $l$ -vector [8], which is a combinatorial characterization of a query point  $q$  with respect to the input set  $P$ . In their study, Durocher et al. [18] introduced the  $\psi$ -histogram and established its direct link to several depth measures for  $q$ , including Tukey depth, simplicial depth and the perihedral depth mentioned above. Depth histograms are another such summarization technique of depths of points in  $P$  based on the combinatorial structure of  $P$  [8]. We briefly discussed these concepts in Section 2.3. We suggest one possible direction for a future study is to extend the idea of  $\psi$ -histogram and depth histograms for a set of query points and use that information for analysis of the spread of points in  $Q$  relative to  $P$ .

**Open problem 6:** Study a possible extension of  $\psi$ -histograms and depth histograms for sets of points.

Depth measures are compared and contrasted using their properties. But the properties introduced in previous literature are designed to describe depth measures defined for a single query point. In Chapter 5, we discussed seven of these properties, and we presented generalizations for five of these properties to the depth of a set of points: P1.1 affine invariance, P2.1 maximality at the center of symmetry, P3.1 monotonicity relative to the deepest point, P4.1 vanishing at infinity, and P5.1 consistency across dimensions. We examined these properties for our newly defined simplicial and Tukey depths of a set and for other generalizations for depth of a set discussed in Chapter 2. We noted that the newly defined Tukey depth of a set of points and new definition for simplicial depth of a set of query points follow P1-P4 similarly to their respective generalizations.

**Open Problem 7:** Study properties of depth measures for sets.

Finally, some questions remain unanswered with respect to improving the running times of the algorithms presented in Theorems 1 and 2. In particular, an important contribution would be to establish respective lower bounds expressed in terms of  $n$  and  $k$  on the running times required to compute the simplicial depth and Tukey depth of a set of  $k$  points relative to a set of  $n$  points in  $\mathbb{R}^d$ .

<p><b>Open Problem 8:</b> Establish lower bounds on the worst-case time required to compute batched depth queries of simplicial and Tukey depth.</p>
--

# Bibliography

- [1] Peyman Afshani, Donald R. Sheehy, and Yannik Stein. Approximating the simplicial depth in high dimensions. In *Proceedings of 32nd European workshop on Computational Geometry*, pages 39–42, 2016.
- [2] Dror Aiger, Haim Kaplan, and Micha Sharir. Duality-based approximation algorithms for depth queries and maximum depth. *arXiv preprint arXiv:2006.12318*, 2020.
- [3] Greg Aloupis. Geometric measures of data depth. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 72:147–158, 2006.
- [4] Greg Aloupis, Carmen Cortés, Francisco Gómez, Michael Soss, and Godfried Toussaint. Lower bounds for computing statistical depth. *Computational Statistics & Data Analysis*, 40(2):223–229, 2002.
- [5] Greg Aloupis, Stefan Langerman, Michael Soss, and Godfried Toussaint. Algorithms for bivariate medians and a Fermat-Torricelli problem for lines. *Computational Geometry*, 26(1):69–79, 2003.
- [6] Luis Barba, Stefan Lochau, Alexander Pilz, and Patrick Schnider. Simplicial depth for multiple query points. In *Proceedings of 35th European Workshop on Computational Geometry*, page 7 pages, 2019.
- [7] Vic Barnett. The ordering of multivariate data. *Journal of the Royal Statistical Society: Series A (General)*, 139(3):318–355, 1976.
- [8] Daniel Bertschinger, Jonas Passweg, and Patrick Schnider. Tukey depth histograms. In *Proceedings of 33rd International Workshop on Combinatorial Algorithms, IWOCA 2022*, pages 186–198, 2022.

- [9] Vladimir G. Boltyanski and J. Jerónimo Castro. Centrally symmetric convex sets. *Journal of Convex Analysis*, 14(2):345, 2007.
- [10] Endre Boros and Zoltán Füredi. The number of triangles covering the center of an  $n$ -set. *Geometriae Dedicata*, 17(1):69–77, 1984.
- [11] David Bremner, Dan Chen, John Iacono, Stefan Langerman, and Pat Morin. Output-sensitive algorithms for tukey depth and related problems. *Statistics and Computing*, 18(3):259–266, 2008.
- [12] Michael A. Burr, Eynat Rafalin, and Diane L. Souvaine Souvaine. Simplicial depth: An improved definition, analysis, and efficiency for the finite sample case. *Series in Discrete Mathematics and Theoretical Computer Science (DIMACS)*, 72:195–209, 2006.
- [13] Timothy M. Chan. An optimal randomized algorithm for maximum tukey depth. In *Proceedings of Annual Symposium on Discrete Algorithms (SODA)*, volume 4, pages 430–436, 2004.
- [14] Bernard Chazelle, Leo J. Guibas, and Der-Tsai Lee. The power of geometric duality. *BIT Numerical Mathematics*, 25(1):76–90, 1985.
- [15] Dan Chen, Pat Morin, and Uli Wagner. Absolute approximation of tukey depth: Theory and experiments. *Computational Geometry*, 46(5):566–573, 2013.
- [16] Andrew Y. Cheng and Ming Ouyang. On algorithms for simplicial depth. In *Proceedings of the 13th Canadian Conference of Computation Geometry (CCCG)*, pages 53–56, 2001.
- [17] David L. Donoho and Miriam Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, 20(4):1803–1827, 1992.
- [18] Stephane Durocher, Robert Fraser, Alexandre Leblanc, Jason Morrison, and Matthew Skala. On combinatorial depth measures. *International Journal of Computational Geometry & Applications*, 28(04):381–398, 2018.

- [19] Herbert Edelsbrunner and Leonidas J. Guibas. Topologically sweeping an arrangement. *Journal of Computer and System Sciences*, 38(1):165–194, 1989.
- [20] Herbert Edelsbrunner and Emo Welzl. Constructing belts in two-dimensional arrangements with applications. *SIAM Journal on Computing*, 15(1):271–284, 1986.
- [21] Anil K. Ghosh and Probal Chaudhuri. On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, 32(2):327–350, 2005.
- [22] Joseph Gil, William Steiger, and Avi Wigderson. Geometric medians. *Discrete Mathematics*, 108(1-3):37–51, 1992.
- [23] Myeong-Hun Jeong, Yaping Cai, Clair J. Sullivan, and Shaowen Wang. Data depth based clustering analysis. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–10, 2016.
- [24] Samir Khuller and Joseph SB Mitchell. On a triangle counting problem. *Information Processing Letters*, 33(6):319–321, 1990.
- [25] David Kirkpatrick. Optimal search in planar subdivisions. *SIAM Journal on Computing*, 12(1):28–35, 1983.
- [26] Stefan Langerman and William Steiger. Optimization in arrangements. In *Proceedings of Symposium on Theoretical Aspects of Computer Science*, pages 50–61. Springer, 2003.
- [27] D. T. Lee and Yu-Tai Ching. The power of geometric duality revisited. *Information Processing Letters*, 21(3):117–122, 1985.
- [28] Jun Li and Regina Y. Liu. New nonparametric tests of multivariate locations and scales using data depth. *Statistical Science*, 19(4):686–696, 2004.
- [29] Regina Y. Liu. On a notion of data depth based on random simplices. *The Annals of Statistics*, pages 405–414, 1990.
- [30] Regina Y. Liu. Control charts for multivariate processes. *Journal of the American Statistical Association*, 90(432):1380–1387, 1995.

- [31] Hendrik P. Lopuhaa and Peter J. Rousseeuw. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, 19(1):229–248, 1991.
- [32] Kim Miller, Suneeta Ramaswami, Peter Rousseeuw, J. Antoni Sellares, Diane Souvaine, Ileana Streinu, and Anja Struyf. Efficient computation of location depth contours by methods of computational geometry. *Statistics and Computing*, 13(2):153–162, 2003.
- [33] Pavlo Mozharovskyi. Tukey depth: linear programming and applications. *arXiv e-prints*, pages arXiv–1603, 2016.
- [34] Hannu Oja. Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, 1(6):327–332, 1983.
- [35] Alexander Pilz and Patrick Schneider. Extending the centerpoint theorem to multiple points. In *Proceedings of 29th International Symposium on Algorithms and Computation (ISAAC 2018)*, volume 123, page 53:1–53:13. Schloss Dagstuhl-Leibniz-Zentrum für Informatik GmbH, 2018.
- [36] Alexander Pilz, Emo Welzl, and Manuel Wettstein. From crossing-free graphs on wheel sets to embracing simplices and polytopes with few vertices. *Discrete & Computational Geometry*, 64(3):1067–1097, 2020.
- [37] Peter J. Rousseeuw and Mia Hubert. Regression depth. *Journal of the American Statistical Association*, 94(446):388–402, 1999.
- [38] Peter J. Rousseeuw and Ida Ruts. Bivariate location depth. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 45(4):516–526, 1996.
- [39] Peter J. Rousseeuw, Ida Ruts, and John W. Tukey. The bagplot: a bivariate boxplot. *The American Statistician*, 53(4):382–387, 1999.
- [40] Peter J. Rousseeuw and Anja Struyf. Computing location depth and regression depth in higher dimensions. *Statistics and Computing*, 8(3):193–203, 1998.
- [41] Ida Ruts and Peter J. Rousseeuw. Computing depth contours of bivariate point clouds. *Computational statistics & data analysis*, 23(1):153–168, 1996.

- [42] Robert Serfling. Depth functions in nonparametric multivariate inference. *Series in Discrete Mathematics and Theoretical Computer Science (DIMACS)*, 72:1, 2006.
- [43] Michael I. Shamos and Dan Hoey. Geometric intersection problems. In *Proceeding of Symposium on Foundations of Computer Science*, pages 208–215. IEEE, 1976.
- [44] John W. Tukey. Mathematics and the picturing of data. In *Proceeding of International Congress of Mathematicians*, volume 2, pages 523–531, 1975.
- [45] Yijun Zuo and Robert Serfling. General notions of statistical depth function. *Annals of statistics*, pages 461–482, 2000.