FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

BACHELOR THESIS

Erik Mendroš

Simplicial depth

Department of Probability and Mathematical Statistics

Supervisor of the bachelor thesis: Mgr. Stanislav Nagy, Ph.D.
Study programme: General Mathematics

Prague 2023

I would like to express my sincere gratitude to my supervisor Mgr. Stanislav Nagy, Ph.D. for his consistent encouragement and guidance throughout my research work. His contagious enthusiasm and genuine interest in the subject has been a constant source of inspiration and motivation for me. I would also like to thank my family for their unwavering support, encouragement, and love.

I declare that I carried out this bachelor thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In .... on .... Author signature

Název práce: Simplexová hĺbka

Autor: Erik Mendroš

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Stanislav Nagy, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Hĺbkové funkcie nepochybne zohrávajú kľúčovú úlohu v neparametrickej štatistike, a to tým, že zovšeobecňujú poradie a kvantily pre viacrozmerné dáta. V našej práci sa zameriame na simplexovú hĺbkovú funkciu. Dôkladne dokážeme jej hlavné vlastnosti, pričom dôkazy doplníme o ilustrácie. Tiež si predstavíme niektoré z možných alternatívnych definícií simplexovej hĺbky. Počas štúdie jednej z nich však narazíme na určité nepresnosti v publikovaných výsledkoch. Pokúsime sa ich opraviť a v niektorých prípadoch aj rozšíriť. Nakoniec, v záverečnej časti práce predstavíme zaujímavú súvislosť medzi simplexovou hĺbkou a Sylvesterovym problémom štyroch bodov, ktorá môže mať dôsledky pre budúce pokroky v tejto oblasti.

Klíčová slova: hĺbka dát, simplexová hĺbka, medián, viacrozmerné dáta

Title: Simplicial depth

Author: Erik Mendroš

Department: Department of Probability and Mathematical Statistics

Supervisor of the bachelor thesis: Mgr. Stanislav Nagy, Ph.D., Department of Probability and Mathematical Statistics

Abstract: Depth functions play a crucial role in nonparametric statistics by generalizing orderings, ranks, and quantiles to multivariate data. In our thesis, we provide a comprehensive study of the classical and revised definitions of simplicial depth function, accompanied by detailed and illustrated proofs of some of their properties. Our research also addresses some issues in previous publications and explores potential expansions of those concepts. In the final part of the thesis, we reveal an intriguing connection between simplicial depth and Sylvester's four-point problem, which may have implications for future advancements in this field.

Keywords: data depth, simplicial depth, median, multivariate data

Contents

Introduction

Over the last few decades, various definitions of the so-called statistical depth function have been introduced. In essence, a statistical depth function is a tool used to determine how “central” a particular point in $\mathbb{R}^d$ is with respect to a probability distribution in the same space. The function assigns greater values to points that are situated nearer to the distribution’s center, while smaller values are assigned to points that are further from the center. The rationale for introducing a depth function is to extend the idea of quantiles beyond one-dimensional data and into multi-dimensional spaces. This extension is driven by the need to identify central tendencies within datasets that have multiple dimensions, similar to how quantiles are used to identify central tendencies in one-dimensional data.

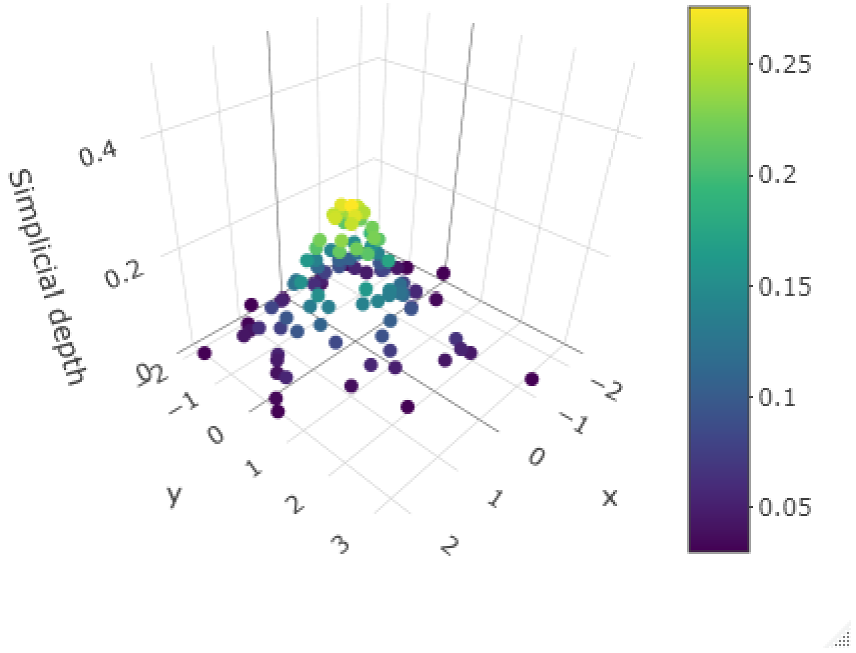


Figure 1: The figure displays a sample of 100 points randomly drawn from a standard bivariate normal distribution. The assigned value of the simplicial depth function for each point is represented by its z-coordinate. Points with higher depth are considered to be more centrally located, whereas those with lower depth are positioned towards the distribution’s periphery. The point with the greatest depth can be viewed as a generalization of a median to higher dimensions.

Bearing in mind the importance of quantiles in nonparametric statistics, it is of no surprise that this generalization has become increasingly pursued over the years. The quantile function in $\mathbb{R}$ relies on the ordering of data, a concept that is relatively straightforward to define in the one-dimensional case due to the natural ordering of real numbers. However, there is no equivalent canonical ordering in the multivariate scenario, which makes the generalization of quantiles for multivariate data more challenging and allows for multiple definitions to be introduced. One such definition is the simplicial depth, to which this work is dedicated.

The simplicial depth function was first introduced in 1988 by Liu [11]. Since then, multiple attempts have been made to alter and revise Liu's original definition of simplicial depth. In Chapter 1 we will discuss and compare a few of them. Naturally, there are some essential properties which are desired for a depth function to possess, and so our assessments will be based on these properties. Since the concept of simplicial depth is fairly geometrical, theorems and definitions will be often times accompanied by illustrations to enhance understanding.

In Chapter 2, our focus shifts towards exploring the characteristics of a revised definition of the simplicial depth function. Specifically, we examine the sample simplicial depth that was initially proposed by Burr, Rafalin and Souvaine in [5]. However, our investigation revealed an error in their work that led to several unpleasant consequences regarding their findings. Thus, our objective is to rectify their reasoning and, in some instances, even expand the theory further.

Lastly, in final Chapter 3 we unveil a relationship between simplicial depth and Sylvester's four-point problem. Consequently, we discuss a minor paradox that we came across during our investigation. This requires us to present our findings on the computation of population simplicial depth. To be more specific, we provide the exact value of the simplicial depth of the centroid of a triangle in $\mathbb{R}^2$, with respect to the uniform continuous distribution on that triangle in the same space. To the best of our knowledge, such result is original.

1 Simplicial depth

The purpose of this chapter is to summarize some of the main definitions that have arisen along the way of fiddling with the simplicial depth function. Each of these definitions is essentially a variation of the original concept with minor alterations. In spite of that, their behavior may differ greatly, as will be demonstrated on examples. Then, after introducing all the necessary terminology, we will provide rigorous proofs for the four primary properties of a depth function.

1.1 General depth function

If not stated otherwise, consider d to be a natural number. All random variables in the thesis are defined on a common probability space $(\Omega,\mathcal{A},\mathsf{P})$. By $\mathcal{P}(\mathbb{R}^d)$ we will understand the class of all distributions on the Borel sets in $\mathbb{R}^d$ and by P_X the distribution of a given random vector X. In the introduction we slightly touched what the general notion of statistical depth means. For further analysis, however, it is important to formulate it more precisely. Several intuitively desirable properties were formulated by Zuo and Serfling [19] in a general definition of depth function. In their work, the attention was confined to depth functions that are nonnegative and bounded. For such functions were then formulated the following four properties:

(P1) *Affine invariance.* The depth of a point $x \in \mathbb{R}^d$ should not depend on the underlying coordinate system.

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

(P3) *Monotonicity relative to the deepest point.* As a point $x \in \mathbb{R}^d$ moves away from the point with the maximal depth function value along any fixed ray through the "deepest" point, the depth at x should decrease monotonically.

(P4) *Vanishing at infinity* As $\|x\|$ approaches infinity, the depth of a point x should approach zero.

Putting all four main properties together, we arrive at Zuo's and Serfling's general notion of a depth function.

Definition 1 (statistical depth function)**.** *Let the mapping $D(\cdot;\cdot) : \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^1$ be bounded, non-negative, that satisfies the following conditions:*

(i) $D(Ax+b; P_{AX+b}) = D(x; P_X)$ *holds for any random vector* X *in* $\mathbb{R}^d$*, any non-singular matrix* $A \in \mathbb{R}^{d\times d}$ *and any vector* $b \in \mathbb{R}^d$*;*

(ii) $D(\theta; P_X) = \sup_{x\in\mathbb{R}^d} D(x; P_X)$ *holds for any* P_X *that is symmetric about* $\theta \in \mathbb{R}^d$*;*

(iii) For any P_X *having the deepest point (point with the highest depth)* θ *and any* $\lambda \in [0,1]$*, we have* $D(x; P_X) \leq D(\lambda x + (1-\lambda)\theta; P_X)$*;*

(iv) $\lim_{\|x\|\to\infty} D(x; P_X) = 0$.

Then $D(\cdot; P_X)$ *is called statistical depth function.*

In Definition 1 we used the term "symmetric about θ." Since multiple notions of multi-dimensional symmetry are possible, we introduce three that are of our highest interest.

Definition 2 (halfspace)**.** *Let* $u \in \mathbb{R}^d$ *be an arbitrary unit vector (that is* $\|u\| = 1$*) and* $c \in \mathbb{R}$ *a constant. The set of points* $H_{u,c} = \{x \in \mathbb{R}^d : \ u^\top x \leq c\}$ *defines a closed halfspace in* $\mathbb{R}^d$*. Its interior is an open halfspace, and its boundary* $\{x : \ u^\top x = c\}$ *defines a hyperplane.*

Definition 3 (multivariate symmetry)**.** *Let* $\theta \in \mathbb{R}^d$ *and let* X *be a random vector in* $\mathbb{R}^d$*. Its distribution is said to be:*

(i) centrally symmetric about θ *if* $\mathsf{P}(X-\theta \in H) = \mathsf{P}(X-\theta \in -H)$*, for all* $H \subset \mathbb{R}^d$*,* H *is a closed halfspace;*

(ii) angularly symmetric about θ *if* $\mathsf{P}(X-\theta \in H) = \mathsf{P}(X-\theta \in -H)$*, for all* $H \subset \mathbb{R}^d$*,* H *is a closed halfspace with the origin on its boundary;*

(iii) halfspace symmetric about θ *if* $\mathsf{P}(X \in H) \geq 1/2$ *for all* $H \subset \mathbb{R}^d$*,* H *is a closed halfspace with* θ *on its boundary.*

The definitions of central and angular symmetry are not in their original form. However, the conditions above are equivalent to the original ones as was shown for central symmetry by Zuo and Serfling in [20, Lemma 2.1] and indirectly[1] for angular symmetry by Rousseeuw and Struyf [16, Theorem 1]. We opt to use these alternative definitions, since it is now straightforward to see the following implications:

$$C\text{-symmetry} \implies A\text{-symmetry} \implies H\text{-symmetry},$$

where C,A,H will abbreviate central, angular and halfspace, respectively. The converse implications do not hold as can be seen in Figure 1.1.

[1] Even though it was not stated explicitly, it follows from their work, which can be seen in [21, Theorem 7].

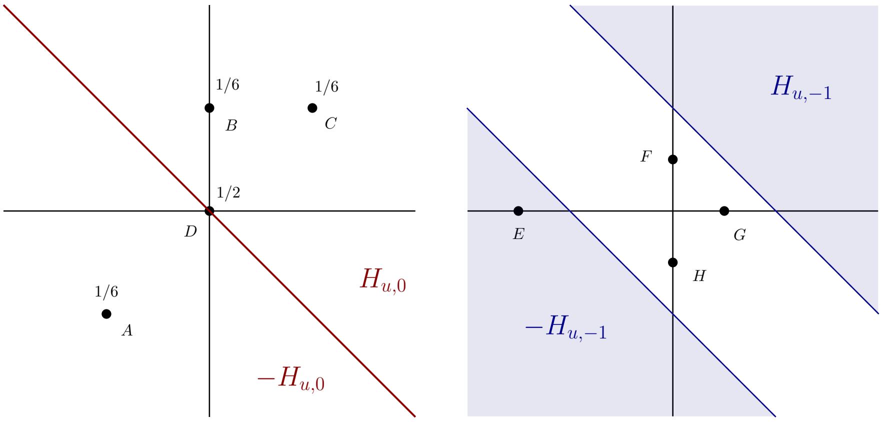


Figure 1.1: H-symmetry, not A-symmetry (left; $d = 2$): consider a discrete probability distribution P_X, whose support is comprised of points A,B,C and D as in the figure, with the corresponding probability displayed above each point. Then, P_X is trivially H-symmetric about the origin D. However, P_X is not A-symmetric; for $u = (-1, -1)^\top$ we have $\mathsf{P}(X \in H_{u,0}) = 5/6$, while $\mathsf{P}(X \in -H_{u,0}) = 4/6$.
A-symmetry, not C-symmetry (right; $d = 2$): take a uniform discrete probability distribution P_X at points E,F,G and H arranged as in the figure. Then P_X is A-symmetric, while not C-symmetric (about the origin). For $u = (-1, -1)^\top$ we have $\mathsf{P}(X \in H_{u,-1}) = 0 \neq 1/4 = \mathsf{P}(X \in -H_{u,-1})$, contradicting the definition of C-symmetry.

1.2 Definitions and terminology

At first, let us start with defining the notion of a simplex upon which the whole concept of simplicial depth stands.

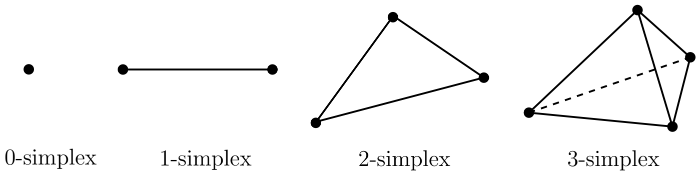
Definition 4 (affine and convex hull)**.** *For $K \subset \mathbb{R}^d, K \neq \emptyset$ we define an affine hull and convex hull as*

$$\operatorname{aff}(K) = \left\{ \sum_{x \in \boldsymbol{I}} \lambda(x)x : \sum_{x \in \boldsymbol{I}} \lambda(x) = 1,\ \emptyset \neq \boldsymbol{I} \subset K,\ \boldsymbol{I} \textit{ is finite} \right\},$$

$$\operatorname{conv}(K) = \left\{ \sum_{x \in \boldsymbol{I}} \lambda(x)x : \lambda(x) \geq 0\ \forall x \in \boldsymbol{I},\ \sum_{x \in \boldsymbol{I}} \lambda(x) = 1,\ \emptyset \neq \boldsymbol{I} \subset K,\ \boldsymbol{I} \textit{ is finite} \right\},$$

respectively.

Definition 5 (simplex)**.** *By k-dimensional simplex (k-simplex for short) S with vertices $x_1, \ldots, x_{k+1} \in \mathbb{R}^d$, where $k \leq d$, we understand the convex hull of the set of points $\{x_1, \ldots, x_{k+1}\}$. We denote it by $S(x_1, \ldots, x_{k+1})$. Additionally, by writing simplex we will always understand a d-simplex.*



Definition 6 (simplicial depth)**.** *The simplicial depth of a point x in $\mathbb{R}^d$ with respect to $P_X \in \mathcal{P}(\mathbb{R}^d)$ is defined to be the probability that x belongs to a random simplex in $\mathbb{R}^d$, that is,*

$$SD(x; P_X) = P(x \in S(X_1, \ldots, X_{d+1})), \tag{1.1}$$

where $X_1, \ldots, X_{d+1}$ is a random sample from P_X.

Definition 6 is the original population notion of the simplicial depth introduced by Liu [11]. Note that with this definition we use the word population. One of the main motivations to define the depth function was to find the multivariate median of some given dataset. The probability measure from which our dataset comes is often unknown and only a sample of points $x_1, \ldots, x_n$ is observed. To help us define the sample version of simplicial depth we represent the distribution of our data by an empirical measure $\hat{P}_n \in \mathcal{P}(\mathbb{R}^d)$. The probability distribution $\hat{P}_n$ corresponds to n (possibly repeated) points $x_1, \ldots, x_n \in \mathbb{R}^d$, with mass $1/n$ attached to each of them.

Definition 7 (sample simplicial depth)**.** *Let $x_1, \ldots, x_n \in \mathbb{R}^d$ be observed data points and let $\hat{P}_n \in \mathcal{P}(\mathbb{R}^d)$ denote the corresponding empirical measure. The sample counterpart to the population simplicial depth is then defined as*

$$SD_n^{closed}(x; \hat{P}_n) = \frac{1}{\binom{n}{d+1}} \sum_{1 \le i_1 < \cdots < i_{d+1} \le n} \mathbb{I}(x \in S(x_{i_1}, \ldots, x_{i_{d+1}})), \tag{1.2}$$

where $\mathbb{I}$ denotes the indicator function.

From here we can see that the sample simplicial depth of a point is basically the proportion of simplices that contain x to the total number of $\binom{n}{d+1}$ simplices, where we only consider simplices generated by distinct sample points (which may nonetheless repeat, if the location of several sample points is the same). The reason for including the term "*closed*" in reference to SD_n will become clearer in what follows later in this chapter.

Notation. Occasionally, we may only be concerned with the relative order of the sample simplicial depth of points. In such scenarios, we can disregard the fraction $1/\binom{n}{d+1}$ in (1.2), as it is the same for every point. To denote this modified depth function, we use a distinct font type:

$$\mathsf{SD}_n^{closed}(x; \hat{P}_n) = \binom{n}{d+1} \cdot SD_n^{closed}(x; \hat{P}_n).$$

If it is clear from the context, we will refer to this modified sample simplicial depth simply as to the sample simplicial depth.

Next, in order to prevent any potential confusion, we introduce a set of supplementary definitions that will help to disambiguate certain concepts later on.

- To a point that is part of a dataset is referred to as a *data point*.
- To a non-dataset point is referred to as a *position*.

- A *convex polytope* of dimension k (or shortly *k-polytope*) is a bounded non-empty intersection of finitely many closed halfspaces in $\mathbb{R}^d$, where the dimension refers to the dimension of the smallest affine subspace of $\mathbb{R}^d$ that contains the entire polytope. One example of a k-polytope is a non-degenerate k-simplex.

- A *face* of a d-polytope D is defined as either D itself, or a non-empty subset of D of the form $D \cap h$, where h is a hyperplane such that D is fully contained in one of the closed halfspaces determined by h. Observe that each face of D is a convex polytope of some dimension less than or equal to d. This is because D is a bounded intersection of finitely many halfspaces and h is the intersection of two halfspaces.

- By a *k-dimensional face* (k-face for short) of a d-polytope D for $0 \leq k \leq d$ we understand a face of D of dimension k, where the dimension of a face is the dimension of the smallest affine subspace that contains it. Notice, that the k-face of a d-simplex is always a k-simplex. For this reason, terms k-face of a d-simplex and k-simplex can be used synonymously.

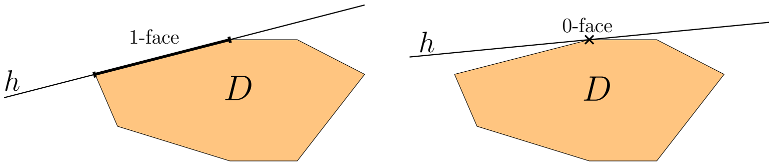


Figure 1.2: The illustrations of a 2-polytope D and its 1-face (left, bold line segment) and 0-face (right, cross).

- A *facet* of a d-polytope D is defined as a $(d-1)$-face of D.

- Let S be a dataset of n points in $\mathbb{R}^d$ in general position,[2] where $n \geq d+1$. All possible $(d-1)$-simplices, defined by d data points from S, subdivide $\mathbb{R}^d$ into regions. And so, last but not least a *cell* is the set of all positions inside the convex hull of S that can be connected by a line segment which does not intersect any $(d-1)$-simplex induced by the observed data points. The closure of a cell classifies as a d-polytope. On the other hand, a cell can be viewed as the interior of some d-polytope. For illustrations see Figure 1.3.

Before we proceed to the revised definitions of the sample simplicial depth, we may want to look at a motivation for why were these other definitions proposed in the first place. Among the problems that arise with Liu's definition of depth function in the finite sample case is the (sometimes) unwanted discontinuity. It is not difficult to see that the depth of all positions on the boundary of a cell is at least the depth of a position in its interior. In fact, it is usually the case that the depth values on the boundaries are higher than the depth in each of the adjacent cells. To demonstrate why this might be a problem we use the examples presented in Figure 1.4 below.

[2] A set of at least $d+1$ points in $\mathbb{R}^d$ is said to lie in general position if no $d+1$ of these points lie on the same hyperplane.

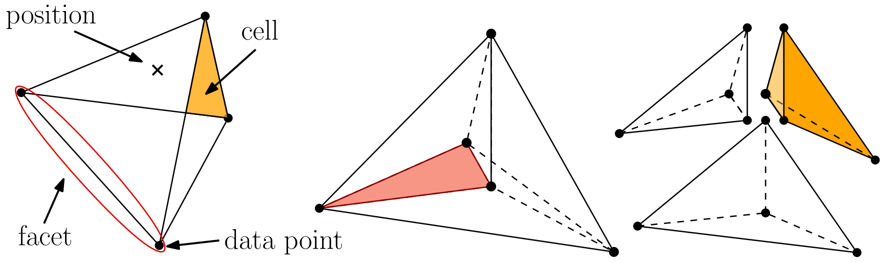


Figure 1.3: Terminology introduced above depicted in 2-dimensional space (left) and 3-dimensional space (middle and right). A facet of a simplex for $d = 2$ is a line segment with data points as endpoints (displayed in a red ellipse), whereas a facet of a simplex for $d = 3$ is a closed triangle with data points as vertices (displayed in red). In these illustrations, a cell for $d = 2$ is an open triangle and a cell for $d = 3$ is an open tetrahedron (both displayed in orange). For clarity, all of the cells in the middle picture were split apart in the right hand panel of the figure.

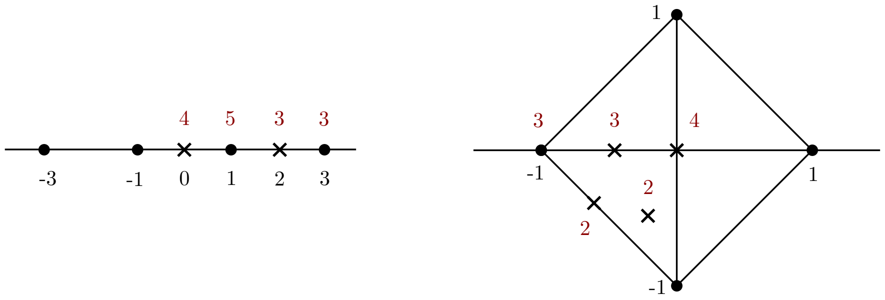


Figure 1.4: (left; $d = 1$): Consider a dataset $\{-3, -1, 1, 3\}$. The corresponding empirical measure $\hat{P}_4$ is C-symmetric about 0 and an easy calculation gives $\mathsf{SD}_4^{closed}(0; \hat{P}_4) = 4$, $\mathsf{SD}_4^{closed}(1; \hat{P}_4) = 5$, violating both **(P2)** and **(P3)**. The sample simplicial depth of points is displayed in red and by the convention below Definition 7 the fraction $1/\binom{4}{2}$ is not included. (right; $d = 2$): Consider a dataset $\{(-1, 0)^\top, (1, 0)^\top, (0, -1)^\top, (0, 1)^\top\}$. Then, the corresponding empirical measure $\hat{P}_4$ is C-symmetric about $(0{,}0)^\top$ and, as can be seen in the figure, both **(P2)** and **(P3)** are satisfied.

After viewing the example depicted on the left in Figure 1.4, one could inquire whether open simplices would not be a more suitable option compared to closed ones. In fact, several authors argue that a depth based on open simplices is to be preferred [6, 7]. Thus, the definition changes to

$$SD_n^{open}(x; \hat{P}_n) = \frac{1}{\binom{n}{d+1}} \sum_{1 \le i_1 < \cdots < i_{d+1} \le n} \mathbb{I}(x \in \operatorname{int}(S(x_{i_1}, \ldots, x_{i_{d+1}}))), \tag{1.3}$$

where by the interior of $S \subset \mathbb{R}^d$, denoted by $\operatorname{int}(S)$, we understand the union of all subsets of S that are open in $\mathbb{R}^d$. Analogously to the notation below Definition 7 we define

$$\mathsf{SD}_n^{open}(x; \hat{P}_n) = \binom{n}{d+1} \cdot SD_n^{open}(x; \hat{P}_n).$$

To see how using open simplices behaves in the counterexamples from Figure 1.4 we refer to Figure 1.5.

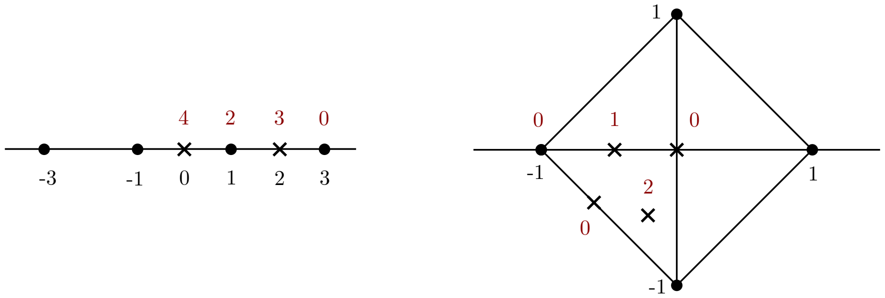


Figure 1.5: This time we used SD_4^{open} to calculate the corresponding depth. Apparently, the use of open simplices fixed property **(P2)** in our counterexample on the left. However, **(P3)** still does not hold. What is worse, the example on the right now violates both **(P2)** and **(P3)**.

Figures 1.4 and 1.5 suggest another solution to this problem. Using closed simplices resulted in the depth of positions lying on the boundary of a cell being greater or equal to those lying inside that cell. Meanwhile, using open simplices did the opposite. That is, the depth of positions inside the cell was greater than or equal to those lying on the boundary of that cell. In [4, 5], Burr, Rafalin and Souvaine came up with an idea to combine both definitions in order to cancel out the above mentioned inconveniences. Their idea was based upon a simple averaging of open and closed simplices:

$$SD_n^{avg}(x; \hat{P}_n) = \frac{1}{2}(SD_n^{closed}(x; \hat{P}_n) + SD_n^{open}(x; \hat{P}_n)); \tag{1.4}$$

$$\mathsf{SD}_n^{avg}(x; \hat{P}_n) = \binom{n}{d+1} \cdot SD_n^{avg}(x; \hat{P}_n).$$

In Figure 1.6, we revisit the counterexamples presented earlier, but this time we use SD_n^{avg}.

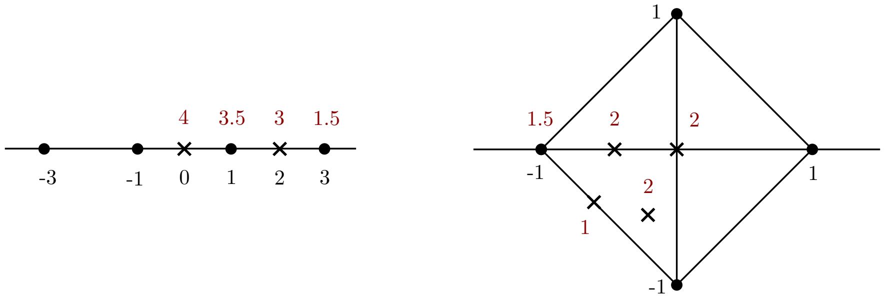


Figure 1.6: Indeed, the behavior of SD_4^{avg} in the given examples seems promising, since **(P2)** and **(P3)** are satisfied in both of them.

To wrap up this section we formulate the following observation.

Observation 1. Consider a set of observed data points and the corresponding empirical measure $\hat{P}_n$. Regardless of the type of the sample simplicial depth used, i.e., SD_n^{closed}, SD_n^{open}, SD_n^{avg}, the sample simplicial depth of any position x within a cell remains the same, that is $SD_n^{closed}(x; \hat{P}_n) = SD_n^{open}(x; \hat{P}_n) = SD_n^{avg}(x; \hat{P}_n)$.

Proof. From the definition of a cell, we have that any position x inside a cell does not intersect any $(d-1)$-simplex induced by the observed data points. Alternatively, we can say that no facet of any simplex contains x. Thus, the simplices containing the position x must contain it in their interior. From that follows $SD_n^{closed}(x;\hat{P}_n) = SD_n^{open}(x;\hat{P}_n)$ and therefore by the definition of SD_n^{avg}, also $SD_n^{closed}(x;\hat{P}_n) = SD_n^{avg}(x;\hat{P}_n)$.

□

1.3 Properties

In this section, we summarize several proven properties that indicate the relevance of using simplicial depth. Our starting point is to verify the consistency of the sample simplicial depth. Afterwards, we prove the four main properties **(P1)**-**(P4)** from Section 1.1 for the population simplicial depth.

1.3.1 Consistency

Probably the most important aspect of any depth function is whether its sample version converges to the population counterpart. In our case, it is indeed possible to prove that the sample simplicial depth is uniformly consistent and to show that, we will help ourselves with known results. However, in our first theorem below, the notion of a random empirical measure is used, and so we need to define it beforehand.

Definition 8 (random empirical measure)**.** *Let $X_1, X_2, \dots, X_n$ be a random sample from $P_X \in \mathcal{P}(\mathbb{R}^d)$. The random empirical measure P_n is a mapping from $\mathcal{B}(\mathbb{R}^d) \times \Omega$ to* [0,1] *defined as*

$$P_n(A;\omega) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_{i(\omega)}}(A), \quad A \in \mathcal{B}(\mathbb{R}^d), \;\; \omega \in \Omega,$$

where δ_y is the Dirac measure at $y \in \mathbb{R}^d$ and $\mathcal{B}(\mathbb{R}^d)$ denotes the Borel σ-algebra on $\mathbb{R}^d$.

Note. The difference between $\hat{P}_n$ and P_n is in randomness. While the empirical measure $\hat{P}_n$ was deterministic (non-random), constructed from one realization of a random sample, the random empirical measure P_n is constructed using the random sample "directly." Apart from this, notice that for each ω from Ω the map $P_n(\cdot;\omega)$ is a probability measure. That follows from the fact that a Dirac measure is a probability measure and the average of a finite number of probability measures is also a probability measure. Further in this subsection, the notation $P_n(\omega)$ is used as a shorthand for $P_n(\cdot;\omega)$.

Theorem 1. *Let $X_1, X_2, \dots, X_n$ be a random sample from $P_X \in \mathcal{P}(\mathbb{R}^d)$ and $P_n \in \mathcal{P}(\mathbb{R}^d)$ the corresponding random empirical measure. Then the following holds:*

$$\mathsf{P}\left(\left\{\omega \in \Omega : \sup_{x\in\mathbb{R}^d} |SD(x;P_n(\omega)) - SD(x;P_X)| \xrightarrow{n\to\infty} 0\right\}\right) = 1.$$

Theorem 2. *The following statement holds true:*

$$\sup_{\hat{P}_n} \sup_{x \in \mathbb{R}^d} |SD_n^{closed}(x; \hat{P}_n) - SD(x; \hat{P}_n)| = \mathcal{O}(1/n),$$

where the first supremum is taken over all empirical measures $\hat{P}_n \in \mathcal{P}(\mathbb{R}^d)$ and $\mathcal{O}$ refers to the "big O notation."[3]

The proof of Theorem 1 can be found in Dümbgen's paper [9, Corollary 1] and Theorem 2 was proven by Nagy in [14, Theorem 1]. Although not explicitly stated by Dümbgen in [9], the following corollary can be easily derived from his work.

Corollary. The sample simplicial depth is uniformly consistent, meaning

$$\mathsf{P}\left(\left\{\omega \in \Omega : \sup_{x \in \mathbb{R}^d} |SD_n^{closed}(x; P_n(\omega)) - SD(x; P_X)| \xrightarrow{n \to \infty} 0\right\}\right) = 1. \tag{1.5}$$

Proof. (Corollary) From Theorem 2 we have

$$\sup_{x \in \mathbb{R}^d} |SD_n^{closed}(x; \hat{P}_n) - SD(x; \hat{P}_n)| \xrightarrow{n \to \infty} 0$$

for all empirical measures $\hat{P}_n \in \mathcal{P}(\mathbb{R}^d)$. Now, as stated earlier, $P_n(\omega)$ is a probability measure for all $\omega \in \Omega$. More specifically, it is an empirical measure corresponding to n points in $\mathbb{R}^d$. Therefore, by Theorem 2, for every fixed $\omega \in \Omega$ and thus a fixed sequence of empirical measures $\{P_n(\omega)\}_{n=1}^{\infty}$ we can write

$$\sup_{x \in \mathbb{R}^d} |SD_n^{closed}(x; P_n(\omega)) - SD(x; P_n(\omega))| \xrightarrow{n \to \infty} 0. \tag{1.6}$$

Using the triangle inequality we obtain that

$$\begin{aligned}\sup_{x \in \mathbb{R}^d} |SD_n^{closed}(x; P_n(\omega)) - SD(x; P_X)| \leq & \sup_{x \in \mathbb{R}^d} |SD_n^{closed}(x; P_n(\omega)) - SD(x; P_n(\omega))| \\ & + \sup_{x \in \mathbb{R}^d} |SD(x; P_n(\omega)) - SD(x; P_X)|\end{aligned}$$

for all $\omega \in \Omega$. And finally, by Theorem 1 and (1.6), the right-hand side of the triangle inequality equation goes to 0 almost surely. This completes the proof. □

1.3.2 Affine invariance

Proving the first of our four main properties in its full generality can be done without great difficulties. The idea of the proof of Theorem 3 was outlined in [12].

Theorem 3. *Let $A \in \mathbb{R}^{d \times d}$ be a non-singular matrix and $b \in \mathbb{R}^d$. Then for all $x \in \mathbb{R}^d$ and any random vector X in $\mathbb{R}^d$ we have*

$$SD(Ax + b; P_{AX+b}) = SD(x; P_X).$$

[3] By $a_n = \mathcal{O}(b_n)$ as $n \to \infty$ we mean that the sequence $\{a_n/b_n\}_{n=1}^{\infty}$ is well defined and bounded.

Proof. A point $x \in \mathbb{R}^d$ is contained in a simplex $S(x_1, \ldots, x_{d+1})$ if x can be written as convex combination of the given simplex vertices. Therefore, by the definition of a convex combination, checking whether or not a point x is contained in a simplex $S(x_1, \ldots, x_{d+1})$ amounts to solving the following system of $d{+}1$ linear equations for a_i, $i \in \{1{,}2{,}...{,}d+1\}$

$$\boxed{\begin{aligned} &x = a_1 x_1 + a_2 x_2 + \cdots + a_{d+1} x_{d+1}, \\ &1 = a_1 + a_2 + \cdots + a_{d+1}, \\ &a_i \geq 0, \quad i \in \{1, 2, \ldots, d+1\}. \end{aligned}} \tag{1.7}$$

Consequently, we claim that the following equivalence holds

$$x \in S(x_1, \ldots, x_{d+1}) \iff Ax + b \in S(Ax_1 + b, \ldots, Ax_{d+1} + b). \tag{1.8}$$

To prove our claim, let us write

$$\begin{aligned} Ax + b &= a_1(Ax_1 + b) + a_2(Ax_2 + b) \cdots + a_{d+1}(Ax_{d+1} + b) \\ &= A(a_1 x_1 + a_2 x_2 + \cdots + a_{d+1} x_{d+1}) + b(a_1 + \cdots + a_{d+1}) \\ &= A(a_1 x_1 + a_2 x_2 + \cdots + a_{d+1} x_{d+1}) + b. \end{aligned}$$

Subtracting b from both sides of the equation and then multiplying by A^{-1} from the left gives

$$x = a_1 x_1 + a_2 x_2 + \cdots + a_{d+1} x_{d+1}.$$

From here we can see that whenever x belongs to $S(x_1, \ldots, x_{d+1})$, meaning that there exists a set of coefficients $\{a_i\}_{i=1}^{d+1}$ satisfying the conditions in the box (1.7), then the same set of coefficients can be used to write $Ax + b$ as a convex combination of the points $Ax_1 + b, \ldots, Ax_{d+1} + b$, meaning that $Ax + b$ belongs to $S(Ax_1 + b, \ldots, Ax_{d+1} + b)$. Thus, we have proven one implication of the equivalence (1.8) and by the same line of reasoning, the converse implication holds true as well. Hence, by the definition of the simplicial depth (1.1) it follows that $SD(Ax + b; P_{AX+b}) = SD(x; P_X)$ and the proof of **(P1)** is therefore complete. □

1.3.3 Maximality at center and monotonicity

Properties **(P2)** and **(P3)**, in contrast to **(P1)**, were only proved for absolutely continuous and angularly symmetric distributions in [12]. However, as mentioned in [14], the proof can be generalized to all distributions P_X satisfying the so called *smoothness condition*

$$P_X(h) = 0, \quad \text{for every } h \subset \mathbb{R}^d,\ h \text{ is a hyperplane}, \tag{1.9}$$

along with A-symmetry. To all distributions P_X satisfying (1.9) is further referred to as *smooth* distributions. Moreover, it is sufficient to require only H-symmetry. Let θ be the center of H-symmetry. Indeed, by the definition of H-symmetry and the assumption of smoothness (1.9), we have

$$\mathsf{P}(X - \theta \in H) = \mathsf{P}(X - \theta \in -H) = \frac{1}{2}, \quad \text{for all halfspaces } H \in \mathbb{R}^d,$$

which by Definition 3 implies A-symmetry.

Theorem 4. *If $P_X \in \mathcal{P}(\mathbb{R}^d)$ is smooth and halfspace symmetric about the origin, then $SD(\alpha x; P_X)$ is a monotone non-increasing in $\alpha \geq 0$ for all $x \in \mathbb{R}^d$.*

Proof. Let $d = 2$ for simplicity. The object of our interest is the difference $SD(x; P_X) - SD(\alpha x; P_X)$, where $\alpha \geq 1$. Only two types of events (see Figure 1.7) contribute to this difference:

$$\begin{aligned} A_{in} &= [\text{arrow from } x \text{ to } \alpha x \text{ enters the random simplex } S(X_1,X_2,X_3)], \\ A_{out} &= [\text{arrow from } x \text{ to } \alpha x \text{ leaves the random simplex } S(X_1,X_2,X_3)]. \end{aligned}$$

Making our argument precise requires additional notation (for illustrations see Figure 1.8). We write $\overline{a,b}$ for the line segment from a to b and $\overleftrightarrow{a,b}$ for the line containing two distinct points $a,b \in \mathbb{R}^d$ (in this proof $d = 2$). The line $\overleftrightarrow{a,b}$ divides $\mathbb{R}^2$ into two halfplanes. If that line does not contain the origin, we call the closed halfplane with the boundary $\overleftrightarrow{a,b}$ which contains the origin the "inner side," denoted by $I(a,b)$. Let x and $\alpha \geq 1$ be fixed. Let us denote by C the set of all possible pairs $(a,b) \in \mathbb{R}^2 \times \mathbb{R}^2$ whose line segment intersects $\overline{x, \alpha x}$, that is

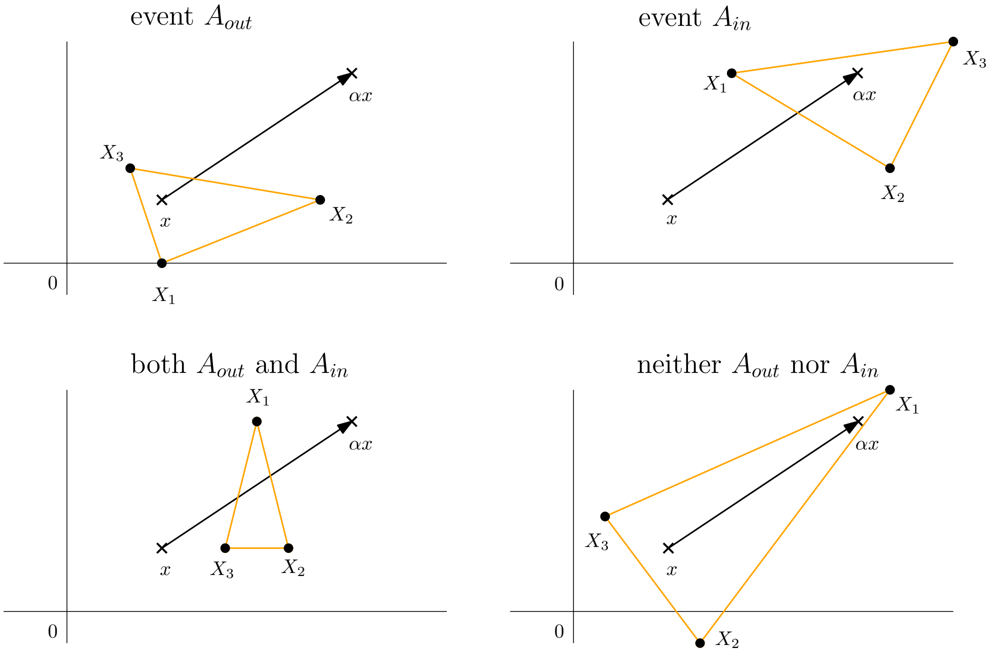


Figure 1.7: Some examples of the notation A_{in} and A_{out}.

$$C = \{(a,b) : \overline{a,b} \cap \overline{x, \alpha x} \neq \emptyset\}.$$

Let us further define events A_{in}^{ij} and A_{out}^{ij} for $i,j,k \in \{1,2,3\}$, $i \neq j \neq k$:

$$\begin{aligned} A_{in}^{ij} &= [\{(X_i, X_j) \in C\} \cap \{X_k \notin \operatorname{int}(I(X_i, X_j))\}], \\ A_{out}^{ij} &= [\{(X_i, X_j) \in C\} \cap \{X_k \in I(X_i, X_j)\}]. \end{aligned}$$

Instances of events $A_{out}^{23}, A_{in}^{12}, A_{in}^{13}$ can be seen in Figure 1.7, more precisely in the subfigures: top left, top right, bottom left, respectively. Clearly, A_{in} is a subset of $A_{in}^{12} \cup A_{in}^{23} \cup A_{in}^{13}$. Only those events, where all three points X_1, X_2, X_3

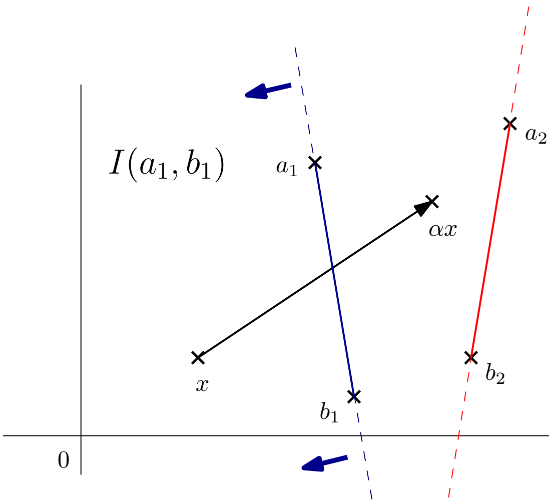


Figure 1.8: To clarify the notation, the pair (a_1, b_1) belongs to the set C, while the pair (a_2, b_2) does not, since the line segment $\overline{a_2{,}b_2}$ does not intersect the line segment $\overline{x{,}\alpha x}$. The line $\overleftrightarrow{a_1{,}b_1}$ divides the plane into two halfplanes, where $I(a_1{,}b_1)$ denotes the one containing the origin (blue arrows).

lie on a hyperplane (a line in this case) containing x, contribute to the union while not to the set A_{in}. Therefore, using the assumption of smoothness, we get $\mathsf{P}(A_{in}) = \mathsf{P}(A_{in}^{12} \cup A_{in}^{23} \cup A_{in}^{13})$. Similar remark holds for the set A_{out}.

Let i,j and k continue to refer to the same values as before. Then,

$$\begin{aligned} A_{in}^{ij} \cap A_{in}^{jk} = [&\{(X_i, X_j) \in C\} \cap \{X_k \notin \operatorname{int} I(X_i, X_j)\} \\ &\cap \{(X_j, X_k) \in C\} \cap \{X_i \notin \operatorname{int} I(X_j, X_k)\}]. \end{aligned}$$

This means that in order for an event to belong to this intersection, the only possible configuration of points X_i,X_j,X_k is for them to lie on a line, with $\overline{X_i, X_j}$ and $\overline{X_j, X_k}$ intersecting $\overline{x{,}\alpha x}$. And so, using the assumption of smoothness and neglecting sets with zero probability, the events A_{in}^{ij} and A_{in}^{jk} are disjoint for all possible combinations of i,j,k.

The three events A_{in}^{ij} are equally probable since the orders of observations X_i, X_j, X_k are also equally probable. The same remarks (disjunction and equal probability) hold for the three events A_{out}^{ij} as well. Now, with $\alpha \geq 1$ and $x \in \mathbb{R}^2$ fixed, the event $A_{out} \setminus A_{in}$ includes those random simplices that contain x but do not contain αx and the event $A_{in} \setminus A_{out}$ includes those random simplices that contain αx but do not contain x. Let us additionally abbreviate the random simplex $S(X_1, X_2, X_3)$ by S. Finally, using all of our notation, we may now rewrite the difference $SD(x; P_X) - SD(\alpha x; P_X)$ in the following manner:

$$
\begin{aligned}
&SD(x; P_X) - SD(\alpha x; P_X) = \\
&= \mathsf{P}(x \in S) - \mathsf{P}(\alpha x \in S) \\
&= \mathsf{P}(x \in S, \alpha x \notin S) + \mathsf{P}(x \in S, \alpha x \in S) - (\mathsf{P}(\alpha x \in S, x \notin S) + \mathsf{P}(\alpha x \in S, x \in S)) \\
&= \mathsf{P}(x \in S, \alpha x \notin S) - \mathsf{P}(\alpha x \in S, x \notin S) \\
&= \mathsf{P}(A_{out} \setminus A_{in}) - \mathsf{P}(A_{in} \setminus A_{out}) \\
&= \mathsf{P}(A_{out}) - \mathsf{P}(A_{out} \cap A_{in}) - (\mathsf{P}(A_{in}) - \mathsf{P}(A_{in} \cap A_{out})) \\
&= \mathsf{P}(A^{12}_{out} \cup A^{23}_{out} \cup A^{13}_{out}) - \mathsf{P}(A^{12}_{in} \cup A^{23}_{in} \cup A^{13}_{in}) \\
&= 3\mathsf{P}(A^{12}_{out}) - 3\mathsf{P}(A^{12}_{in}) \\
&= 3 \int_{(x_1,x_2)\in C} \mathsf{P}(X_3 \in I(x_1, x_2)) - \mathsf{P}(X_3 \notin \operatorname{int} I(x_1, x_2)) \,\mathrm{d}\, P_X(x_1) \,\mathrm{d}\, P_X(x_2) \\
&= 3 \int_{(x_1,x_2)\in C} (2\mathsf{P}(X_3 \in I(x_1, x_2)) - 1) \,\mathrm{d}\, P_X(x_1) \,\mathrm{d}\, P_X(x_2),
\end{aligned}
$$

where in the last equality we used the assumption of smoothness (1.9) in order to write

$$
\mathsf{P}(X_3 \notin \operatorname{int} I(x_1, x_2)) = \mathsf{P}(X_3 \notin I(x_1,x_2)) = 1 - \mathsf{P}(X_3 \in I(x_1,x_2)).
$$

The last integral is greater than or equal to 0 thanks to the assumption of H-symmetry which gives $\mathsf{P}(X_3 \in H) \geq 1/2$, for any closed halfspace H with origin on its boundary. Indeed, there exists $H \subseteq I(a,b)$, for all $a,b \in \mathbb{R}^2, a \neq b$ and for such H we obtain $\mathsf{P}(X_3 \in I(x_1,x_2)) \geq \mathsf{P}(X_3 \in H) \geq 1/2$. This proves the assertion.

□

Note that in the original version of Theorem 4, the assumption of absolute continuity of the distribution was used when neglecting null sets in order to split A_{in} and A_{out} into $A^{12}_{in}, A^{23}_{in}, A^{13}_{in}$ and $A^{12}_{out}, A^{23}_{out}, A^{13}_{out}$, respectively, and to show that these triplets are disjoint. As was seen in the proof, however, replacing the condition of the absolute continuity by the condition of smoothness preserved the correctness of the proof.

Moving on to property **(P2)**, we start by proving an auxiliary lemma, whose proof was not included in Liu's papers.

Lemma 5. *Let $X_1, X_2, \ldots, X_{d+1}$ be random vectors from smooth distribution $P_X \in \mathcal{P}(\mathbb{R}^d)$ and for $i = 1, 2, \ldots, d+1$ denote $X_i^* = X_i / \|X_i\|$. Then, except for zero probability sets, the following four events are equivalent*

(i) $\{(X_1, \ldots, X_{d+1}) : \ 0 \in S(X_1, \ldots, X_{d+1})\}$;

(ii) $\{(X_1, \ldots, X_{d+1}) : \ 0 \in S(X_1^*, \ldots, X_{d+1}^*)\}$;

(iii) $\{(X_1, \ldots, X_{d+1}) : \ 0 \in S(e_1, \ldots, e_d, V)\}$, *where e_i is the i-th canonical vector in $\mathbb{R}^d$ and $V = [X_1^*| \cdots |X_d^*]^{-1} X_{d+1}^*$, where $[X_1^*| \cdots |X_d^*]$ is the matrix with columns $X_1^*, \ldots, X_d^*$;*

(iv) $\{(X_1, \ldots, X_{d+1}) : \ V^{(1)} < 0, \ldots, V^{(d)} < 0\}$; *where $V^{(i)}$ is the i-th component of the vector V.*

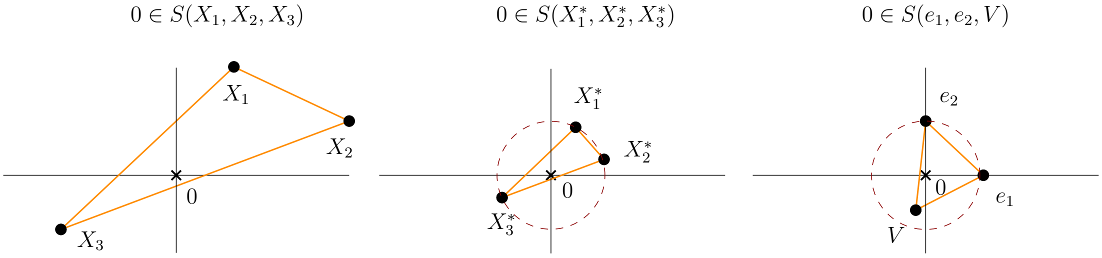


Figure 1.9: To gain an intuitive understanding of the equivalences, we provide illustrations of the first three events in Lemma 5.

Proof. $(i) \iff (ii)$
It is sufficient to establish the implication

$$0 \in S(X_1, \dots, X_{d+1}) \implies 0 \in S(cX_1, X_2, \dots, X_{d+1}), \text{ for any } c > 0.$$

The actual equivalence from the lemma then follows by induction and by setting c appropriately. Note that due to the assumption of smoothness (1.9) we have $\mathsf{P}(\|X_1\| \neq 0) = 1$, and so we only consider events where all the points $X_1, \dots, X_{d+1}$ are non-zero. Assume that $0 \in S(X_1, \dots, X_{d+1})$. Then, for $X_1, \dots, X_{d+1}$ as vertices, there exists a set of coefficients $\{a_i\}_{i=1}^{d+1}$ fulfilling the conditions in the box (1.7). We aim to find a set of coefficients $\{b_i\}_{i=1}^{d+1}$ that satisfies these conditions for $cX_1, X_2, \dots, X_{d+1}$. Let us set

$$\overline{b_1} = a_1, \quad \overline{b_i} = c \cdot a_i, \quad i = 2, \dots, d+1.$$

Now the sum $\sum_{i=1}^{d+1} \overline{b_i}$ is equal to some $K > 0$. We claim that by setting $b_i = \overline{b_i}/K$ for $i = 1, \dots, d+1$ we obtain the correct set of coefficients. We indeed have $b_i \geq 0$ and $\sum_{i=1}^{d+1} b_i = 1$. To prove that 0 can be expressed as a convex combination of points $cX_1, X_2, \dots, X_{d+1}$ with coefficients $\{b_i\}_{i=1}^{d+1}$, we use the assumption $0 \in S(X_1, \dots, X_{d+1})$ and write

$$\begin{aligned} 0 &= \sum_{i=1}^{d+1} a_i X_i \\ 0 &= \sum_{i=1}^{d+1} \frac{a_i c}{K} X_i \\ 0 &= b_1 c X_1 + \sum_{i=2}^{d+1} b_i X_i. \end{aligned}$$

$(ii) \iff (iii)$
This equivalence follows from (1.8) with $A = [X_1^* | \cdots | X_d^*]^{-1}$ and $b = 0 \in \mathbb{R}^d$, where again, assuming smoothness (1.9) gives $\mathsf{P}(A \text{ is non-singular}) = 1$.

$\neg(iv) \implies \neg(iii)$
We aim to prove that when at least one $V^{(i)} > 0$, then 0 is not contained within the simplex $S(e_1, \dots, e_d, V)$. Note that we do not need to consider the case where $V^{(i)} = 0$ for any $i = 1, \dots, d$, since from smoothness (1.9) we have $\mathsf{P}(\bigcup_{i=1}^{d} [V^{(i)} = 0]) = 0$. Without loss of generality, let $V^{(1)} > 0$. Take a convex combination

$$W = a_1 e_1 + \cdots + a_d e_d + a_{d+1} V. \tag{1.10}$$

In order for the first element of W to be zero, the coefficients a_1 and a_{d+1} must be zero. Furthermore, in order for the j-th element ($j = 2,\dots,d$) of W to be zero, the coefficient a_j must be zero. Therefore, in order for W to be $0 \in \mathbb{R}^d$, we obtain $a_i = 0$ for all $i = 1,\dots,d+1$, which results in a contradiction with $\{a_i\}_{i=1}^{d+1}$ being coefficients of convex combination as $\sum_{i=1}^{d+1} a_i = 0 \neq 1$.

$(iv) \implies (iii)$

Set $a_i = -V^{(i)}$ for $i = 1,\dots,d$ and $a_{d+1} = 1$. It is not difficult to see that

$$a_1 e_1 + \cdots + a_d e_d + a_{d+1} V = 0.$$

Now, $\sum_{i=1}^{d+1} a_i = K > 0$. By dividing all of the coefficients a_i by K, we get

$$a_i/K \geq 0, \quad \sum_{i=1}^{d+1} a_i/K = 1, \quad \frac{a_1}{K} e_1 + \cdots + \frac{a_d}{K} e_d + \frac{a_{d+1}}{K} V = 0,$$

which implies the statement given in (iii).

□

Theorem 6. *If $P_X \in \mathcal{P}(\mathbb{R}^d)$ is smooth and halfspace symmetric about $\theta \in \mathbb{R}^d$, then $SD(\theta; P_X) = 2^{-d}$.*

Proof. Using the already proven affine invariance in Theorem 3, we might without loss of generality assume that P_X is H-symmetric about the origin. As stated at the beginning of this subsection, under the conditions of this theorem H-symmetry implies A-symmetry. With the use of the original definition[4] of A-symmetry (as listed in [11]) we know that X_i^* and $-X_i^*$ are identically distributed, where $X_i^* = X_i / \|X_i\|$. Then, as we proved in Lemma 5, the following two events are (except for sets with zero probability) equivalent:

- $\{(X_1,\dots,X_{d+1}) : \ 0 \in S(X_1,\dots,X_{d+1})\}$;
- $\{(X_1,\dots,X_{d+1}) : \ V^{(1)} < 0,\dots,V^{(d)} < 0\}$; where $V^{(i)}$ is the ith component of the vector $V = [X_1^* | \cdots | X_d^*]^{-1} X_{d+1}^*$.

In the next step we demonstrate that by changing X_i^* to $-X_i^*$, the random vector $(V^{(1)},\dots,V^{(d)})^\top$ changes to $(V^{(1)},\dots,V^{(i-1)},-V^{(i)},V^{(i+1)},\dots,V^{(d)})^\top$. Without loss of generality let $i = 1$, then

$$\begin{aligned}
[-X_1^* | X_2^* | \cdots | X_d^*]^{-1} X_{d+1}^* &= [[X_1^* | \cdots | X_d^*] \operatorname{diag}(-1,1,\dots,1)]^{-1} X_{d+1}^* \\
&= \operatorname{diag}(-1,1,\dots,1)^{-1} [X_1^* | \cdots | X_d^*]^{-1} X_{d+1}^* \\
&= \operatorname{diag}(-1,1,\dots,1) V,
\end{aligned}$$

where diag is a diagonal $d \times d$ matrix. Thus,

$$[-X_1^* | X_2^* | \cdots | X_d^*]^{-1} X_{d+1}^* = \begin{pmatrix} -V^{(1)} \\ V^{(2)} \\ \vdots \\ V^{(d)} \end{pmatrix}$$

[4] A random variable $X \in \mathbb{R}^d$ or its distribution P_X is said to be angularly symmetric about $\theta \in \mathbb{R}^d$ if and only if $(X - \theta)/\|X - \theta\|$ and $-(X - \theta)/\|X - \theta\|$ are identically distributed. In case of $P_X(\{\theta\}) > 0$ we use the convention $0/0 = 0$.

from which we obtain that the probability of $(V_1, \dots, V_d)^\top$ is the same as the probability of the same random vector with the signs of its coordinates arbitrary changed. This further implies that each orthant[5] has an equal probability which must be 2^{-d}, since there are precisely 2^d possible orthants. Therefore,

$$\begin{aligned} SD(0; P_X) &= \mathsf{P}(\{(X_1, \dots, X_{d+1}) :\ 0 \in S(X_1, \dots, X_{d+1})\}) \\ &= \mathsf{P}(\{(X_1, \dots, X_{d+1}) :\ V_1 < 0, \dots, V_d < 0\}) \\ &= 2^{-d}. \end{aligned}$$

□

Theorems 4 and 6 yield the following.

Corollary. Given a smooth and halfspace symmetric distribution P_X, it follows that $SD(x; P_X)$ is at most 2^{-d} for any $x \in \mathbb{R}^d$.

It is worth noting that the tight upper bound for the simplicial depth of any absolutely continuous probability distribution in $\mathbb{R}^d$ is still an open problem for $d \geq 2$.

1.3.4 Vanishing at infinity

To conclude this section, we establish the last of the main properties, whose proof is once again sourced from [12]. Note that property **(P4)** "vanishing at infinity" is proven in its full generality as it was with the affine invariance **(P1)**.

Theorem 7. *For any $P_X \in \mathcal{P}(\mathbb{R}^d)$, we have that* $\sup_{\|x\| \geq n} SD(x; P_X) \xrightarrow{n \to \infty} 0.$

Proof. Given $x \in \mathbb{R}^d$, we observe that the event $[x \in S(X_1, \dots, X_{d+1})]$ is contained in the event $[\exists i \in \{1, \dots, d+1\} :\ \|X_i\| \geq \|x\|]$. This can be shown using the triangle inequality for the norm and a fact that whenever x belongs to a simplex S, it can be written as a convex combination of its vertices. Formally,

$$\|x\| = \|a_1 X_1 + \cdots + a_{d+1} X_{d+1}\| \leq a_1 \|X_1\| + \cdots + a_{d+1} \|X_{d+1}\| \leq \max_{i=1,\dots,d+1} \|X_i\|,$$

for some $a_i \geq 0$, $i \in \{1, \dots, d+1\}$, $\sum_{i=1}^{d+1} a_i = 1$. Thus, we obtain

$$[x \in S(X_1, \dots, X_{d+1})] \subset \left[\bigcup_{i=1}^{d+1} (\|X_i\| \geq \|x\|) \right].$$

Let us denote by A_n the event $\left[\bigcup_{i=1}^{d+1} (\|X_i\| \geq n)\right]$, where $n \in \mathbb{N}$. Then we have $A_1 \supset A_2 \supset \cdots$ and $\lim_{n \to \infty} A_n = \bigcap_n A_n = \emptyset$. For each x such that $\|x\| \geq n$, $n \in \mathbb{N}$ we can write

$$\begin{aligned} \mathsf{P}(x \in S(X_1, \dots, X_{d+1})) &\leq \mathsf{P}\left(\bigcup_{i=1}^{d+1} (\|X_i\| \geq \|x\|) \right) \\ &\leq \mathsf{P}\left(\bigcup_{i=1}^{d+1} (\|X_i\| \geq n) \right) \xrightarrow{n \to \infty} \mathsf{P}(\emptyset) = 0. \end{aligned}$$

[5] An orthant is a region of space that is defined by the signs of its coordinates. In two dimensions, an orthant is one of the four quadrants of the coordinate plane.

The convergence in the second line of the equation follows from the fact that a probability measure is continuous from above [8, Theorem 3.1.1]. Since it holds for each x with norm greater than or equal to n, it also holds for the supremum. The claim follows.

□

2 Averaged simplicial depth

The main objective of this chapter is to examine the averaged simplicial depth (1.4). As was suggested in Figure 1.6, the averaged simplicial depth indeed alleviated some of the problems that arose with Liu's original definition. The revised definition appeared to smooth out the irregularities created at boundaries which caused property **(P3)** not to be satisfied. As a result, the concept of simplicial depth could now seem more appealing for statistical analysis. However, despite these improvements, there are still counterexamples where **(P3)** or even **(P2)** is not satisfied. Furthermore, there may be cases where the use of SD_n^{closed} from (1.2) could yield better results.

In the following counterexample to **(P3)**, we will use an averaged analogue of the population simplicial depth (1.1) which we define under the circumstances of Definition 6 as

$$SD^{avg}(x; P_X) = \frac{1}{2}\left(\mathsf{P}(x \in S(X_1, \ldots, X_{d+1})) + \mathsf{P}(x \in \operatorname{int} S(X_1, \ldots, X_{d+1}))\right). \tag{2.1}$$

Example 2.1 (Violation of **(P3)**). To show that SD^{avg} does not fix the problem with the monotonicity property **(P3)**, we borrow a counterexample presented in Zuo's and Serfling's work [19, Counterexample 3]. Originally, the counterexample was used to show that Liu's simplicial depth fails to be maximized at the center of H-symmetry. However, their calculation of depth of points (as we will see later) was inaccurate, leading to false conclusions.

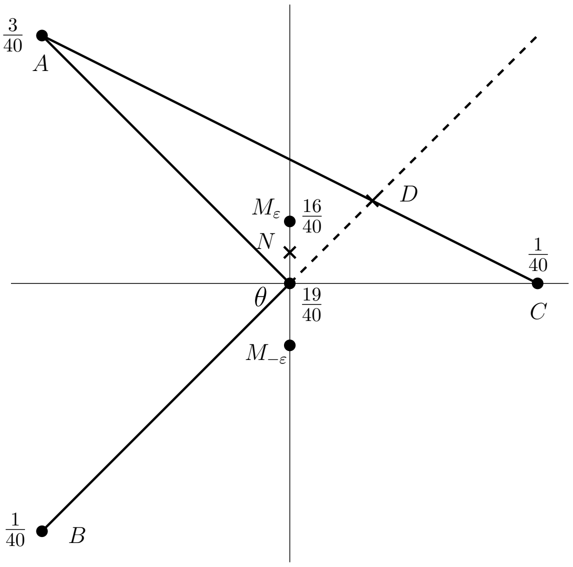


Figure 2.1: Illustration of the distribution described in Example 2.1.

Let $d = 2$ and let $P_X \in \mathcal{P}(\mathbb{R}^2)$ be a distribution (see Figure 2.1) such that $P_X(\{\theta\}) = 19/40$, $P_X(\{A\}) = 3/40$, and $P_X(\{B\}) = P_X(\{C\}) = 1/40$, where $\theta = (0{,}0)^\top$, $A = (-1{,}1)^\top$, $B = (-1, -1)^\top$, $C = (1{,}0)^\top$. The last point is to be taken arbitrary from the open triangle with vertices θ, A and D, where $D = (\frac{1}{3}, \frac{1}{3})^\top$ is the intersection point of lines $\overleftrightarrow{B,\theta}$ and $\overleftrightarrow{A,C}$. So, for any $0 < \varepsilon < 1/2$, we set our last point as $M_\varepsilon = (0, \varepsilon)^\top$, with $P_X(\{M_\varepsilon\}) = 16/40$. The distribution P_X is H-symmetric about θ. Besides, by calculating all of the possible simplicial depth values it can be shown that the maximum simplicial depth of P_X is attained only at θ. These calculations are straightforward but lengthy and similar to those given below; we do not include them. Choose and denote any point inside the line segment $\overline{\theta,M_\varepsilon}$ by N. In order for **(P3)** to be fulfilled, the simplicial depth of N must be greater than or equal to the simplicial depth of M_ε, while, at the same time, less than or equal to the simplicial depth of θ. The calculation of simplicial depth of these three points using Liu's depth (1.1) gives

$$
\begin{aligned}
SD(\theta; P_X) \;&=\; \underbrace{\frac{3!}{40^3}(3+16)}_{3!(P(S(A,B,C))+P(S(M_\varepsilon,B,C)))} + \underbrace{\frac{3!\cdot 19}{40^3}(1+3+3+16+16+3\cdot 16)}_{3!P(S(\theta,Y,Z),\ Y\neq Z,\ Y,Z\in\{A,B,C,M_\varepsilon\})} \\
&\quad + \underbrace{\frac{3\cdot 19}{40^3}\left(1+1+3^2+16^2\right)}_{3P(S(\theta,Y,Y),\ Y\in\{A,B,C,M_\varepsilon\})} + \underbrace{\frac{3\cdot 19^2}{40^3}(1+1+3+16)}_{3P(S(\theta,\theta,Y),\ Y\in\{A,B,C,M_\varepsilon\})} + \underbrace{\frac{19^3}{40^3}}_{P(S(\theta,\theta,\theta))} \\
&= \frac{54853}{40^3}, \\
SD(N; P_X) &= \frac{3!}{40^3}(3+16+3\cdot 19) + \frac{3!\cdot 16\cdot 19}{40^3}(1+3+3) \\
&\quad + \frac{3\cdot 16\cdot 19^2}{40^3} + \frac{3\cdot 16^2\cdot 19}{40^3} \\
&= \frac{41496}{40^3}, \\
SD(M_\varepsilon; P_X) &= \frac{3!}{40^3}(3+3\cdot 19) + \frac{3!\cdot 16}{40^3}(1+3+3+19+19+3\cdot 19) \\
&\quad + \frac{3\cdot 16}{40^3}\left(1+1+3^2+19^2\right) + \frac{3\cdot 16^2}{40^3}(1+1+3+19) + \frac{16^3}{40^3} \\
&= \frac{50536}{40^3}.
\end{aligned}
$$

The mistake in [19] was due to not considering degenerate simplices, which resulted in M_ε having greater SD than θ, contradicting **(P2)**. Despite that not being the case, it is clear that the simplicial depth of N is less than the simplicial depth of M_ε, and so our depth function is not monotone on that particular ray.

The first summands in each of the formulae above correspond to the triangles that contain the reference point in their interior. Halving all the terms of the equations, except those first summands, allows us to derive SD^{avg} for these three points. This fact becomes clearer with the upcoming observation. We thus have

$$
SD^{avg}(\theta; P_X) = \frac{27483.5}{40^3}, \quad SD^{avg}(N; P_X) = \frac{20976}{40^3}, \quad SD^{avg}(M_\varepsilon; P_X) = \frac{25448}{40^3},
$$

which means that property **(P3)** remains violated even for SD^{avg}. △

In spite of SD^{avg} not being successful in this case, we can observe that the difference was a little less extreme compared to SD. This raises another question: could we potentially employ a weighted average, rather than an unweighted one, to scale down the depth of point M_ε to a level where monotonicity is satisfied? To further elaborate upon this idea, it is beneficial to introduce the rephrased definitions presented in [4, 5]. We formulate them as an observation.

Observation 2. The averaged population simplicial depth SD^{avg} from (2.1) can be rewritten as

$$SD^{avg}(x; P_X) = \mathsf{P}(x \in \operatorname{int} S(X_1, \dots, X_{d+1})) + \frac{1}{2}\mathsf{P}(x \in \operatorname{bd} S(X_1, \dots, X_{d+1})),$$

where the symbol bd stands for the boundary of a set. Similarly, the averaged sample simplicial depth SD_n^{avg} from (1.4) can be expressed as

$$SD_n^{avg}(x; \hat{P}_n) = \frac{1}{\binom{n}{d+1}} \left(\rho(x; \hat{P}_n) + \frac{1}{2}\sigma(x; \hat{P}_n) \right), \tag{2.2}$$

where $\rho(x; \hat{P}_n)$ is the number of data-determined simplices which contain x in their interior, and $\sigma(x; \hat{P}_n)$ is the number of data-determined simplices which contain x in their boundary.

With the use of Observation 2, we can now reformulate the problem as finding a fitting constant $c \in (0,1)$ so that the depth function

$$SD^{avg,c}(x; P_X) = \mathsf{P}(x \in \operatorname{int} S(X_1, \dots, X_{d+1})) + c \cdot \mathsf{P}(x \in \operatorname{bd} S(X_1, \dots, X_{d+1}))$$

fulfills **(P3)**. Let us explore this idea further using our example.

Example 2.2 (Example 2.1 continued). In our specific scenario, we are basically looking for c such that

$$\frac{50176}{40^3} \cdot c + \frac{360}{40^3} < \frac{41040}{40^3} \cdot c + \frac{456}{40^3} < \frac{54739}{40^3} \cdot c + \frac{114}{40^3}.$$

Unfortunately, a constant $c \in (0,1)$ that would satisfy both of the inequalities does not exist. This means that even a weighted average is not sufficient to scale down the depth of these points appropriately.

It is also important to point out, that the constant c would depend on the geometry of the points. Changing M_ε to $M_{-\varepsilon}$ (with the mass $\frac{16}{40}$) and consequently N to lie inside $\overline{\theta, M_{-\varepsilon}}$ would result in a system of inequalities

$$\frac{50176}{40^3} \cdot c + \frac{132}{40^3} < \frac{41040}{40^3} \cdot c + \frac{420}{40^3} < \frac{54739}{40^3} \cdot c + \frac{306}{40^3}.$$

In this altered case, any $c \in (\frac{6}{721}, \frac{18}{571})$ would satisfy both of the inequalities, meaning that $SD^{avg,c}$ would be monotonous on that part of the ray. △

2.1 Properties of averaged simplicial depth

As we saw in the previous example, the simplicial depth function applied to finitely atomic[6] distributions appeared to be problematic. Aside, we may ask, whether or not it even makes sense to scale down the depth of the atoms of the distribution. For this reason, our focus in this section will be exclusively on probability distributions that are smooth (in the sense of (1.9)). For such distributions and any $x \in \mathbb{R}^d$, the probability of x lying on the boundary of a random simplex would always amount to zero. This means that SD^{avg} in (2.1) would reduce to Liu's original simplicial depth SD from (1.1). Thus, we shall only work with its sample counterpart SD_n^{avg} from (1.4) and consider only data points sampled from smooth probability distributions. Nevertheless, for all the assertions stated in this section, the condition of smoothness may be relaxed to a requirement of points lying in general position.

The most essential properties/advantages of averaged sample simplicial depth were formulated in [4, 5]. Nonetheless, the proof of their first proposition was based on an incorrect argumentation. We demonstrate this in a counterexample, which can be also considered an example of SD_n^{closed} providing more plausible results when compared to SD_n^{avg}. Similarly, the two propositions [4, 5, Corollary 1 and Proposition 2] that followed the first (incorrect) one, were consequently also inaccurate, as their proof was based on the first proposition. We will therefore try to reformulate and partly extend the original assertions from [4, 5]. As we will be relying mostly on the averaged sample simplicial depth, we will at times abbreviate it and call it a *depth* (in this chapter only). For brevity, we will write $SD_n^{avg}(x)$ with the second argument in the depth function omitted. We begin by stating the following lemma, borrowed from [4, 5, Lemma 1], which will prove useful in the proofs presented in this section.

Lemma 8. *For any version of the sample simplicial depth ((1.2), (1.3), (1.4)), the depth of any two positions in the same cell is equal.*

Proof. Let A be a cell defined by d-dimensional data points $x_1, \dots, x_n$ and let y and z be two positions in A. The conclusion drawn in Observation 1 states that all variants of the sample simplicial depth are identical for positions inside any cell. By the definition of a cell, the line segment $L = \overline{y,z}$ lies entirely within A and no $(d-1)$-simplex defined by d data points intersects L.

Assume that y does not have the same sample simplicial depth as z. Without loss of generality, let the depth of y be greater than the depth of z. Consequently there exists a simplex which contains the point y but not z. Therefore, the segment L must then intersect the boundary of this simplex (its $(d-1)$-face), giving a contradiction as no $(d-1)$-simplex can intersect the segment L.

□

[6] A measure $P_X \in \mathcal{P}(\mathbb{R}^d)$ is said to be finitely atomic if the support of P_X is finite, meaning that there exists an integer $m \geq 1$ and points $x_1, \dots, x_m$ such that $P_X(\{x_1, \dots, x_m\}) = 1$. We call each such x_i with $P_X(\{x_i\}) > 0$ an atom of P_X.

2.1.1 A counterexample

For contradictions to be seen more clearly, we paraphrase the propositions from [4, 5] prior to our counterexample. While we have adjusted the phrasing to match our own terminology, the intended meaning remains unchanged. In addition, the last proposition requires us to define the term *opposite cells*. However, it is crucial to point out that the original definition of opposite cells was not precise enough and will be corrected along with the propositions in Section 2.1.2. Lastly, the term sample *simplicial median* refers to a point or a set of points with the greatest depth.

Proposition 9 (incorrect)**.** *Let $x_1, \dots, x_n \in \mathbb{R}^d$ be observed data points in general position. The averaged sample simplicial depth of a position on a facet between two cells is equal to the average of the depths of positions in the two adjacent cells.*

Proposition 10 (incorrect)**.** *Let $x_1, \dots, x_n \in \mathbb{R}^d$ be observed data points in general position. The maximum averaged sample simplicial depth is attained in a cell or at a data point.*

Definition 9 (opposite cells, incorrect)**.** *Two cells whose boundaries both contain a position θ lying on the intersection of two or more hyperplanes induced by the observed data points are opposite cells if and only if the two cells lie on opposite sides of every facet that contains θ. It can be shown that every cell has a unique opposite.*

Proposition 11 (incorrect)**.** *Let $x_1, \dots, x_n \in \mathbb{R}^d$ be observed data points in general position. The averaged sample simplicial depth of the position at the intersection between two or more facets is equal to the average of the depths of two opposite cells of the intersection point.*

Example 2.3 (Counterexample for SD_n^{avg}). For any integer $d \geq 3$ consider the following dataset of $d + 2$ points:

$$O = (0,0,\dots,0)^\top \in \mathbb{R}^d, V = (1,1,\dots,1)^\top \in \mathbb{R}^d, D_i = e_i, \text{ for all } i = 1,2,\dots,d,$$

where e_i denotes i-th canonical vector in $\mathbb{R}^d$.

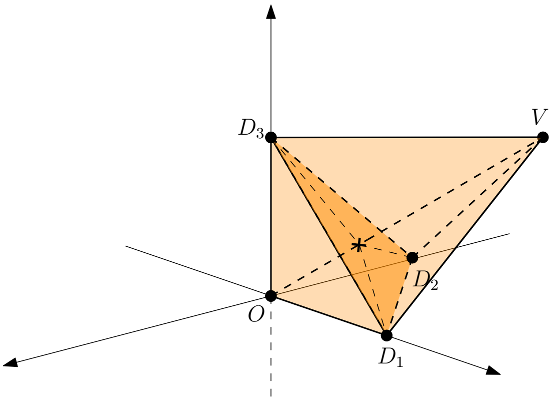
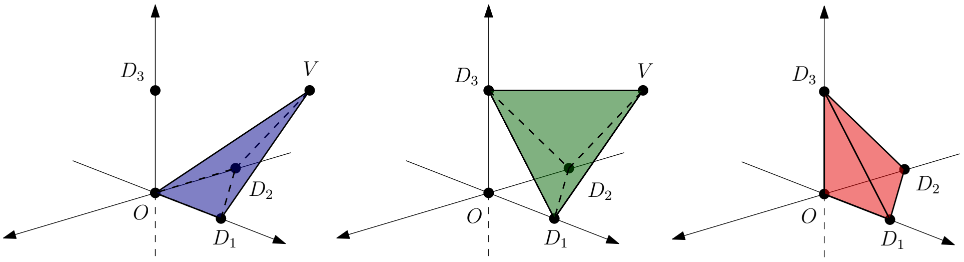


Figure 2.2: The illustration for the situation $d = 3$.

Clearly, such points lie in general position. Thus, by Proposition 9 we should get that SD_{d+2}^{avg} of a position on a facet between two cells is equal to the average of the depths of positions in the two adjacent cells; and by Proposition 10 that the sample simplicial median is attained either in the interior of a cell or at a data point. Neither of that happens to be the case in our setup.

For simplicity, let $d = 3$ first. We have $\binom{5}{4} = 5$ simplices and 6 cells. There are precisely 3 simplices with vertices O, V to which will be referred to as *vertical* simplices. As for the remaining two simplices, one of them has V as a vertex while not O and the other has O as a vertex, while not V. We will call them *upper* and *lower* simplex, respectively.



Vertical simplex

Upper simplex

Lower simplex

As a reminder, we note that the use of different fonts (SD vs. SD) indicates whether or not the depth is normalized by the factor $1/\binom{n}{d+1}$. Calculating the depth of any position inside an arbitrary cell is straightforward, since it must be contained in exactly one interior of a vertical simplex and also in the interior of either the upper or the lower simplex. This means that SD_5^{avg} of any point in any cell is the same and equal to 2. The depth SD_5^{avg} of all data points is also the same, and takes value 2. This follows from the geometry of data points, as each one of them is part of the boundary of the convex hull of all data points, thus cannot be contained in any open simplices. We get

$$\mathsf{SD}_5^{avg}(\text{data point}) = \frac{1}{2} \cdot \binom{4}{3} = 2.$$

Denote $X = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})^\top$ and $Y = (\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon})^\top$ for $\boldsymbol{\varepsilon} \in (0, \frac{1}{3})$. Both X and Y lie on the line segment $\overline{O,V}$, which is part of the boundary of every vertical simplex. Position Y is also contained in the lower simplex and position X was chosen so that it lies on the boundary of the upper and the lower simplex. That leads to

$$\mathsf{SD}_5^{avg}(X) = \frac{1}{2} \cdot 5 = 2.5; \qquad \mathsf{SD}_5^{avg}(Y) = 1 + \frac{1}{2} \cdot 3 = 2.5.$$

Our computations have the following consequences:

- The depth of the positions X and Y is not equal to the average depth of their two adjacent cells, as it is strictly greater than the depth of any cell.
- The sample simplicial median is attained neither in a cell nor at a data point, but in a 1-face instead.
- It is not possible for the depths of both X and Y to be equal to the average depth of two opposite cells. Additionally, it is unclear which cells are considered "opposite" in this context.

Another noteworthy observation is that by the depth calculated above, any point near O or near V that lies on the line segment $\overline{O,V}$ is considered a simplicial median. This version of simplicial median is not an optimal representation of the "hypothetical center" of all data points. Let us therefore compare it with SD_5^{closed} in Figure 2.3.

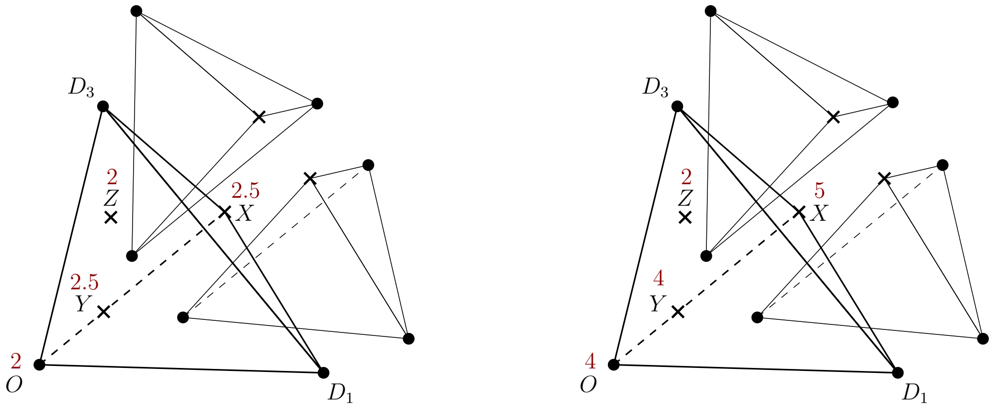


Figure 2.3: The three "bottom" cells (split apart) from Figure 2.2. Depth evaluated using SD_5^{avg} (left), and depth evaluated using SD_5^{closed} (right). In both pictures, the point Z represents any position within any cell.

As can be drawn from the picture, X is the only point at which the simplicial median is attained when considering the depth SD_5^{closed}. This might be viewed as a more desirable outcome, especially when looking only for the simplicial median.

In order to define simplicial median in the case of SD_5^{avg} more appropriately we may have the following idea: define the simplicial median to be a barycenter[7] of M, where M is the set of all points with the greatest depth. Nevertheless, such an idea would consequently make our median sensitive to extreme values in the data. Changing V to $k \cdot V$ in our previous example, where $k > 0$ is some large constant, would result in the "simplicial barycenter median" to be at $\frac{k}{2} \cdot V$, which is far from all the points near the origin. On the other hand, the simplicial median for SD_5^{closed} would stay unchanged at X.

In an analogous manner we now derive all of the above, but for a general dimension $d \geq 3$. Together there are $d+2$ simplices of which d are vertical. Again there is 1 upper and 1 lower simplex. We have $2 \cdot d$ cells and they all have the same depth. This time, we set $X = (\frac{1}{d}, \frac{1}{d}, \ldots, \frac{1}{d})^\top \in \mathbb{R}^d$ and $Y = (\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}, \ldots, \boldsymbol{\varepsilon})^\top \in \mathbb{R}^d$, for any $\boldsymbol{\varepsilon} \in (0, \frac{1}{d})$. Formulae for calculating SD_{d+2}^{avg} of points of our interest are displayed in Table 2.1.

To get formulae for $\mathsf{SD}_{d+2}^{closed}$ we can simply omit all the fractions $\frac{1}{2}$ in Table 2.1. As $\mathsf{SD}_{d+2}^{avg}(X) = \frac{1}{2}(d+2)$ is strictly greater than $\mathsf{SD}_{d+2}^{avg}(Z) = 2$ and $\mathsf{SD}_{d+2}^{avg}(O) = \frac{1}{2}(d+1)$, we arrive at the same contradictions with Propositions 9, 10 and 11 as before. △

[7] The barycenter of a compact set K in $\mathbb{R}^d$ is the expectation of the uniform distribution on K.

point \ dimension	3	4	5	$\cdots$	d
X	2.5	3	3.5	$\cdots$	$\frac{1}{2}(d+2)$
Y	2.5	3	3.5	$\cdots$	$1+\frac{1}{2}d$
Z	2	2	2	$\cdots$	2
O	2	2.5	3	$\cdots$	$\frac{1}{2}(d+1)$

Table 2.1: SD_{d+2}^{avg} of points in different dimensions.

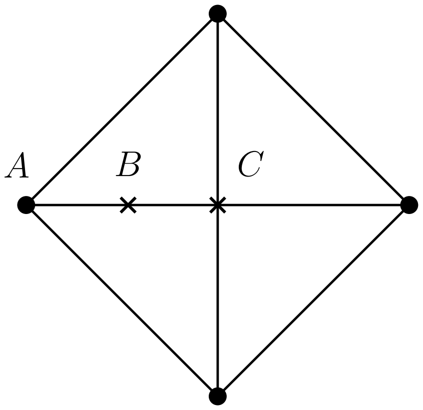
2.1.2 Reformulations and extensions

To lay ground for the reformulated theorems, we begin this subsection with an extended version of the definition from (2.2) and a corrected version of Definition 9. In addition, it is essential to introduce the concept of *relative interior*. The *relative interior* of a set $S \subset \mathbb{R}^d$ (denoted by relint S) is defined as its interior within the smallest affine subspace of $\mathbb{R}^d$ that contains S.

Definition 10. *In the situation from Definition 7, the averaged sample simplicial depth can be expressed as*

$$SD_n^{avg}(x;\hat{P}_n) = \frac{1}{\binom{n}{d+1}}\left(\rho(x;\hat{P}_n) + \frac{1}{2}\sigma_{d-1}(x;\hat{P}_n) + \cdots + \frac{1}{2}\sigma_0(x;\hat{P}_n)\right),$$

where $\rho(x;\hat{P}_n)$ is the number of data-determined simplices which contain x in their interior, and $\sigma_j(x;\hat{P}_n)$ for $j \in \{0,1,\ldots,d-1\}$ is the number of data-determined simplices which contain x in the relative interior of some of their j-face. For brevity, we further omit the second argument of ρ and σ_j. Clearly, $\sigma(x)$ from formula (2.2) is equal to the sum $\sum_{j=0}^{d-1}\sigma_j(x)$.



$$\mathsf{SD}_4^{avg}(x) = \rho(x) + \tfrac{1}{2}\sigma_1(x) + \tfrac{1}{2}\sigma_0(x)$$

$$\mathsf{SD}_4^{avg}(A) = 0 + \tfrac{1}{2}\cdot 0 + \tfrac{1}{2}\cdot 3$$
$$\mathsf{SD}_4^{avg}(B) = 1 + \tfrac{1}{2}\cdot 2 + \tfrac{1}{2}\cdot 0$$
$$\mathsf{SD}_4^{avg}(C) = 0 + \tfrac{1}{2}\cdot 4 + \tfrac{1}{2}\cdot 0$$

Figure 2.4: Definition 10 in practice.

To shed some light on the upcoming definition of opposite cells and to compare it with the original one, we demonstrate its application in Figure 2.5 in the setup of Example 2.3.

Definition 11 (opposite cells - corrected)**.** *Let $x_1, \ldots ,x_n \in \mathbb{R}^d$ be observed data points in general position. Let θ be a position that satisfies the following two conditions:*

(C1) *the position θ lies in at least one $(d-1)$-simplex induced by the observed data points.*

(C2) *the position θ does not lie in any k-simplex induced by the observed data points for all $k < d-1$.*

Two cells whose boundaries both contain the position θ are called opposite cells around θ if and only if the two cells lie on opposite sides of every $(d-1)$-simplex that contains θ.

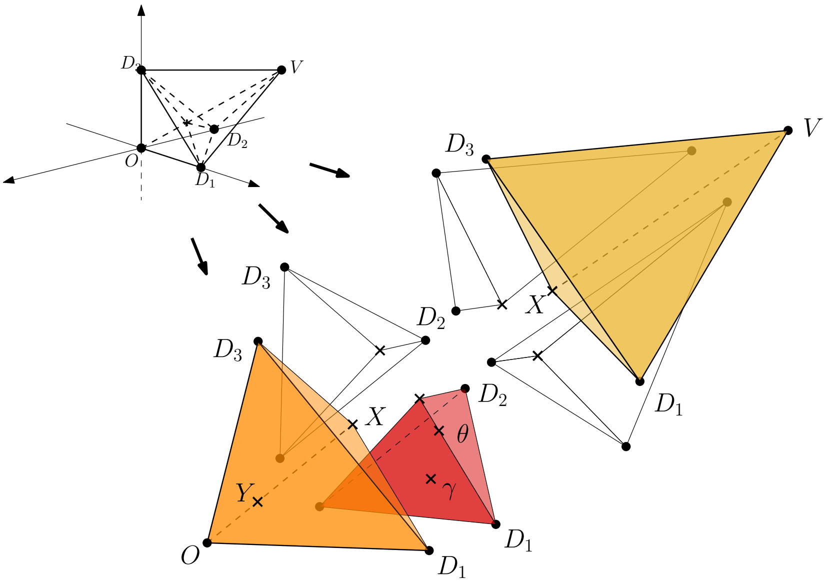


Figure 2.5: For a dataset of 5 points in $\mathbb{R}^3$ we have $\binom{5}{3}$ $(d-1)$-simplices which partition the convex hull of the dataset into six cells. Consider first the positions Y and X which lie at the intersection of three and four $(d-1)$-simplices (facets in the original definition), respectively. According to Definition 9, the orange cell whose boundary contains both X and Y should have a uniquely defined opposite cell. Nevertheless, such a cell in this case does not exist. According to our definition, none of these points satisfy ***(C2)*** as they lie on $(d-2)$-simplex $S(O,V)$. Therefore, the term opposite cell around X and around Y is not defined. Consider now the positions θ and γ. The position θ fulfills ***(C1)***, as it lies at the intersection of $(d-1)$-simplices $S(D_1,D_2,D_3)$ and $S(D_1,O,V)$, but also the condition ***(C2)***. Since the boundary of the red cell contains θ, the opposite cell to the red cell around θ is well defined. As the red cell lies below $S(D_1,D_2,D_3)$, the opposite cell must lie above. As the red cell lies behind $S(D_1,O,V)$, the opposite cell must lie in front of it, which results in the uniquely defined opposite cell depicted in yellow. Lastly, the position γ lies exactly in one $(d-1)$-simplex $S(D_1,O,V)$, thus fulfills ***(C1)*** as well. Since also ***(C2)*** is fulfilled, the opposite (adjacent) cell to the red cell around the position γ is well defined and depicted in orange.

Note. The notion of opposite cells is now correctly defined. First of all, notice that the definition of opposite cell depends on the considered position θ. For two different positions in the boundary of some cell, the opposite cells around these positions may differ. Secondly, to ensure the accuracy of the definition of opposite cells, we had to impose an additional condition ***(C2)*** on the considered position. Lastly, the term *adjacent cells* in Proposition 9 can be viewed as a special case scenario in our definition of opposite cells. That is, two cells are adjacent around θ if θ lies in the boundary of both of these cells and θ lies in the relative interior of precisely one data-determined $(d-1)$-simplex.

Theorem 12 (Propositions 9 and 11)**.** *Let $x_1, \ldots, x_n \in \mathbb{R}^d$ be observed data points in general position. Let θ be a position that satisfies conditions* ***(C1)*** *and* ***(C2)****. Then, the depth of θ is equal to the average of the depths of any pair of opposite cells around θ.*

Proof. Consider any two opposite cells around θ and denote them by A and B. Let x_A be a position in cell A and x_B be a position in cell B. As both x_A and x_B are positions in cells, they can be only contained in an interior of any simplex determined by data points. Hence, we can write

$$\begin{aligned}\mathsf{SD}_n^{avg}(x_A) &= \rho(x_A),\\ \mathsf{SD}_n^{avg}(x_B) &= \rho(x_B),\\ \mathsf{SD}_n^{avg}(\theta) &= \rho(\theta) + \frac{1}{2}\sigma(\theta).\end{aligned}$$

A simplex with θ in its interior: For any data-determined simplex S which contains θ in its interior, there exists $\varepsilon > 0$ such that the ball $B(\theta, \varepsilon)$ is also contained by S. Since θ lies on the boundary of cells A and B, the ball $B(\theta, \varepsilon)$ must contain some points from both of these cells. By Lemma 8, this implies that the simplex S contains cells A and B, thus contributes to $\rho(x_A)$ and $\rho(x_B)$.
A simplex with θ in its boundary: Consider now a data-determined simplex T which contains θ in its boundary. The $(d-1)$-simplex from ***(C1)***, which includes θ must be some $(d-1)$-face of T. Therefore, the interior of the simplex T lies to one side of this facet and the exterior to the other. By the definition of opposite cells, exactly one of A or B lies inside T.

Now, let $\rho^{x_A}(\theta)$ be the number of simplices which contain both θ and x_A in their interiors and let $\sigma^{x_A}(\theta)$ be the number of simplices which contain θ in their boundary and x_A in their interior. Define $\rho^{x_B}(\theta)$ and $\sigma^{x_B}(\theta)$ similarly. By the above argument, we have

$$\rho^{x_A}(\theta) = \rho^{x_B}(\theta) = \rho(\theta).$$

Finally, consider a simplex U, which contains cell A. Then, the point θ must be also contained in the simplex U (either on a boundary or in interior), since for all $\varepsilon > 0$ the ball $B(\theta, \varepsilon)$ contains some points of cell A and by definition, a simplex is a closed set. Thus,

$$\begin{aligned}\mathsf{SD}_n^{avg}(x_A) &= \rho^{x_A}(\theta) + \sigma^{x_A}(\theta),\\ \mathsf{SD}_n^{avg}(x_B) &= \rho^{x_B}(\theta) + \sigma^{x_B}(\theta).\end{aligned}$$

The depth of θ is therefore

$$\begin{aligned}\mathsf{SD}_n^{avg}(\theta) &= \frac{1}{2}(2\rho(\theta)+\sigma(\theta))\\ &= \frac{1}{2}(\rho^{x_A}(\theta)+\rho^{x_B}(\theta)+\sigma^{x_A}(\theta)+\sigma^{x_B}(\theta))\\ &= \frac{1}{2}(\mathsf{SD}_n^{avg}(x_A)+\mathsf{SD}_n^{avg}(x_B)).\end{aligned}$$

□

Note. The requirement of cells A and B to be opposite around θ in Theorem 12 was crucial. Otherwise, the underlined statement in the proof would not hold. In the original proof of Proposition 9, the same statement was made. In their case, however, any position in the shared facet of two cells could be considered, which consequently made the claim of Proposition 9 not true for general d (see Figure 2.6).

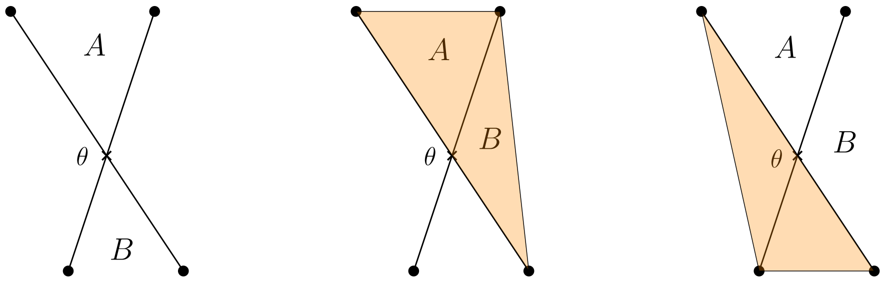


Figure 2.6: If the cells A and B are opposite around θ (left) then every simplex which contains θ in its boundary, contains exactly one of the cells A and B. Whereas, if we only require that the cells share the same facet, there can be simplices with θ in its boundary that contain both of the cells A and B (middle), as well as none of them (right).

For $d = 2$, their assertion in Proposition 10 remains valid since the situation is simplified. For this reason we formulate it individually in the following theorem.

Theorem 13 (Proposition 10, $d = 2$)**.** *Let $x_1, \ldots, x_n \in \mathbb{R}^2$, $n > 3$ be observed data points in general position. Let H denote the convex hull formed by these points. Then the following statements hold true:*

(i) There either exists a data point, or a position in a cell, where the maximum averaged sample simplicial depth is attained.

(ii) The maximum averaged sample simplicial depth is never attained at the boundary of H.

Proof. (i) As $d = 2$, every position satisfies ***(C2)***. Moreover, there are only two "types" of position: a position satisfying ***(C1)***, and a position in a cell. Whenever the maximum sample simplicial depth is attained in a position satisfying ***(C1)***, Theorem 12 guarantees that it is also attained at some cell.

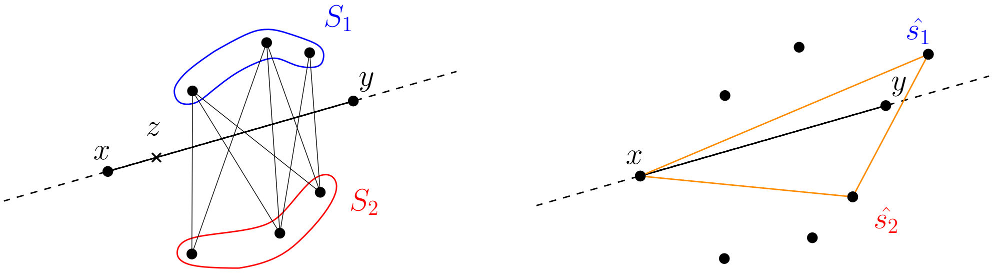
(ii) The depth SD_n^{avg} of any position from the boundary of H equals $\frac{1}{2}(n-2)$, while SD_n^{avg} of any data point from the boundary of H (boundary data points for short) equals $\frac{1}{2}\binom{n-1}{2}$. Since the latter is always greater than the former for all $n > 3$, the maximum sample simplicial depth is not attained at any position from the boundary of H. Now, let $k \leq n$ be the number of boundary data points.

If $k < n$, then there exists an index i such that x_i is contained in the interior of H. By the Carathéodory theorem [18, Theorem 2.2.4], at least one simplex with boundary data points as vertices contains x_i in its interior. Hence,

$$\mathsf{SD}_n^{avg}(x_i) \geq \underbrace{1}_{\rho(x_i)} + \frac{1}{2}\cdot \underbrace{0}_{\sigma_1(x_i)} + \frac{1}{2}\underbrace{\binom{n-1}{2}}_{\sigma_0(x_i)} > \underbrace{0}_{\rho(x_j)} + \frac{1}{2}\cdot \underbrace{0}_{\sigma_1(x_j)} + \frac{1}{2}\underbrace{\binom{n-1}{2}}_{\sigma_0(x_j)} = \mathsf{SD}_n^{avg}(x_j),$$

where x_j is an arbitrary boundary data point. Note that σ_1 on both sides is zero thanks to the assumption of general position.

If $k = n$, then every observed data point is a boundary data point and we need to find a position in the interior of H with SD_n^{avg} greater than $\frac{1}{2}\binom{n-1}{2}$. Let n be even. Take an arbitrary data point x. For this x there always exists some data point y such that half of the remaining data points lie on one side of the line $\overleftrightarrow{x,y}$, while the other half lie on the other side. We denote one of these groups of data points by S_1 and the other one by S_2. We have that $|S_1| = |S_2| = \frac{n-2}{2}$, where $|S|$ is the cardinality of S. Any line segment $\overline{s_1,s_2}$, where $s_1 \in S_1$, $s_2 \in S_2$, must intersect the line segment $\overline{x,y}$. Otherwise, for a pair $\hat{s_1}, \hat{s_2}$ that would not intersect $\overline{x,y}$ we would have that either x lies inside the simplex $S(y,\hat{s_1},\hat{s_2})$ or y lies inside the simplex $S(x,\hat{s_1},\hat{s_2})$. This results in a contradiction with $k = n$. Lastly, denote by z any position lying on $\overline{x,y}$, that is in between x and the closest intersection point of line segments $\overline{x,y}$ and $\overline{s_1,s_2}$, for any $s_1 \in S_1$, $s_2 \in S_2$. Position z is on the boundary of every possible simplex having two of the vertices x and y. At the same time, it is inside of every simplex $S(x, s_1, s_2)$, for any $s_1 \in S_1$, $s_2 \in S_2$. The corresponding depth is thus



$$\mathsf{SD}_n^{avg}(z) = \left(\frac{n-2}{2}\right)^2 + \frac{1}{2}\binom{n-2}{1} = \frac{n^2-2n}{4} > \frac{1}{2}\binom{n-1}{2},$$

where the last inequality holds for $n > 2$.

If n is odd, for an arbitrary data point x we find a data point y (not uniquely defined) so that $|S_1| = \frac{n-1}{2}$ and $|S_2| = \frac{n-3}{2}$. In an analogous manner we arrive at

$$\mathsf{SD}_n^{avg}(z) = \left(\frac{n-1}{2}\right)\cdot\left(\frac{n-3}{2}\right) + \frac{1}{2}\binom{n-2}{1} = \frac{n^2-2n-1}{4} > \frac{1}{2}\binom{n-1}{2},$$

where the last inequality holds for $n > 3$. This proves our assertion.

□

The set of propositions from [4, 5] aimed to simplify the search for the maximum value of SD_n^{avg} by reducing the number/types of points that need to be considered. If Proposition 10 were true for all dimensions d, we could limit our attention to data points and positions within the cells. In other words, when looking for the maximum depth, we could omit all the calculations of depth of positions lying on the boundaries of cells. Unfortunately, we had to impose an additional condition ***(C2)*** on positions for which the statements in Propositions 9 and 11 hold true. As a result, in dimensions $d \geq 3$, there is no relationship between the positions lying in data-determined k-simplices for $k < d-1$ and the positions in the cells. Hence, these positions cannot be excluded when searching for the maximum value of SD_n^{avg}.

Although Proposition 10 does not hold as it was initially thought, we were able to establish a relationship between certain types of positions located in data-determined k-simplices, where $k < d-1$. This relationship will be detailed in Theorem 17 (addition to Theorem 12), and its implications will help us reinforce the correct version of Proposition 10. In order to formulate a follow up to Theorem 12 however, we need a few additional lemmas. We start by formulating an extended version of Lemma 8 for the faces of the closure of a cell.

Lemma 14. *Choose any version of sample simplicial depth ((1.2), (1.3), (1.4)). Then, for $k = 1, \ldots, d-1$, the chosen sample simplicial depth of any two positions in the relative interior of the same k-face of the closure of a cell is the same.*

Proof. The closure of a cell is a d-polytope. Any k-face of a d-polytope is a k-polytope, thus convex. Moreover, the relative interior of a convex set is a convex set. Denote by F the relative interior of a k-face and let x_1 and x_2 be two positions in F. From the convexity of F, the line segment $L = \overline{x_1,x_2}$ lies entirely within F. Any $(d-1)$-simplex that intersects the line segment L must contain the whole line segment. Otherwise, the $(d-1)$-simplex which intersects L but does not contain it entirely, must also intersect the neighboring cells. That contradicts the definition of a cell. The proof can be now carried out in a similar manner as the proof of Lemma 8. Finally, the version of the sample simplicial depth used does not impact the argumentation above.

□

For the next lemma we delve into the intersection theory. Consider a set of observed data points in $\mathbb{R}^3$ in general position. Then, the intersection of any two different data-determined 1-simplices (line segments) is an empty set. Otherwise, the observed data points would not lie in a general position (the endpoints of the line segments would lie in the same hyperplane). In Lemma 15, we show that this observation can be extended to higher dimensions.

Lemma 15. *Let $x_1, \ldots, x_n \in \mathbb{R}^d$, $d \geq 3$ be observed data points in general position. Let U be a data-determined k-simplex, T be a data-determined m-simplex and let $y_1, \ldots, y_\ell$ be their common vertices, where $\ell \in \{0, \ldots, \min(k,m)\}$ and $k + m - \ell \leq d - 1$. Then $U \cap T = \operatorname{conv}(\{y_1, \ldots, y_\ell\})$.*

Proof. It is clear that $\operatorname{conv}(\{y_1, \ldots, y_\ell\}) \subseteq U \cap T$. Suppose, for the sake of contradiction that $\operatorname{conv}(\{y_1, \ldots, y_\ell\}) \subsetneq U \cap T$.

Take an arbitrary point z from the difference $(U \cap T) \setminus \operatorname{conv}(\{y_1, \dots, y_\ell\})$. Since z lies in the intersection of the k-simplex $U = S(y_1, \dots, y_\ell, z_1, \dots, z_{k+1-\ell})$ and the m-simplex $T = S(y_1, \dots, y_\ell, w_1, \dots, w_{m+1-\ell})$, it can be written as two different convex combinations

$$z = \lambda_1 y_1 + \dots + \lambda_\ell y_\ell + \lambda_{\ell+1} z_1 + \dots + \lambda_{k+1} z_{k+1-\ell},$$
$$z = \mu_1 y_1 + \dots + \mu_\ell y_\ell + \mu_{\ell+1} w_1 + \dots + \mu_{m+1} w_{m+1-\ell},$$

where

$$\lambda_t \geq 0 \;\text{ for all } t = 1, \dots, k+1, \quad \sum_{t=1}^{k+1} \lambda_t = 1, \;\text{ and}$$
$$\mu_s \geq 0 \;\text{ for all } s = 1, \dots, m+1, \quad \sum_{s=1}^{m+1} \mu_s = 1.$$

As $z \notin \operatorname{conv}(\{y_1, \dots, y_\ell\})$, at least one μ_t for $t = \ell+1, \dots, m+1$ is greater than zero. Without loss of generality let $\mu_{\ell+1} > 0$. Then w_1 can be expressed as an affine combination of the other vertices as follows

$$\begin{aligned} w_1 = {} & \frac{\lambda_1 - \mu_1}{\mu_{\ell+1}} y_1 + \dots + \frac{\lambda_\ell - \mu_\ell}{\mu_{\ell+1}} y_\ell + \frac{\lambda_{\ell+1}}{\mu_{\ell+1}} z_1 + \dots + \frac{\lambda_{k+1}}{\mu_{\ell+1}} z_{k+1-\ell} \\ & - \frac{\mu_{\ell+2}}{\mu_{\ell+1}} w_2 - \dots - \frac{\mu_{m+1}}{\mu_{\ell+1}} w_{m+1-\ell}. \end{aligned} \tag{2.3}$$

The coefficients of each term in expression (2.3) must sum to 1 in order for it to be classified as an affine combination. This criterion is met, which can be verified by the following computation

$$\sum_{t=1}^{k+1} \frac{\lambda_t}{\mu_{\ell+1}} - \sum_{s=1, s \neq \ell+1}^{m+1} \frac{\mu_s}{\mu_{\ell+1}} = \frac{1}{\mu_{\ell+1}} - \frac{\mu_2}{\mu_{\ell+1}} - \dots - \frac{\mu_{m+1}}{\mu_{\ell+1}} = \frac{\mu_{\ell+1}}{\mu_{\ell+1}} = 1.$$

Finally, denote by M the set of all vertices in U and T. Then M is a set of $k + m - \ell + 2 \leq d + 1$ points and $\operatorname{conv} M = \operatorname{conv}(M \setminus \{w_1\})$. In order for points $x_1, \dots, x_n$ to be in a general position, no N of them lie in a $(N-2)$-dimensional subspace, where $N = 2, 3, \dots, d+1$. In other words, any subset of N points must span an $(N-1)$-dimensional subspace. However the set $M \setminus \{w_1\}$, a set of $k + m - \ell + 1$ points, can span at most a $(k + m - \ell)$-dimensional subspace. Consequently, we arrive at a contradiction with points $x_1, \dots, x_n$ being in a general position, as set M can span at most a $(k + m - \ell)$-dimensional subspace.

□

The subsequent lemma is essentially a corollary of the previous lemma, phrased in terms of Definition 10.

Lemma 16. *Let $x_1, \dots, x_n \in \mathbb{R}^d$, $d \geq 3$ be observed data points in general position. Suppose we have a point $x \in \mathbb{R}^d$ with the depth*

$$SD_n^{avg}(x) = \frac{1}{\binom{n}{d+1}} \left(\rho(x) + \frac{1}{2} \sigma_{d-1}(x) + \dots + \frac{1}{2} \sigma_0(x) \right),$$

where $\sigma_k(x)$ is nonzero for some fixed $k = 0, 1, 2, \dots, d-1$. Then $\sigma_m(x) = 0$ for all $m \in \{0, \dots, d-k-1\}, m \neq k$. If $m = k$ and $m \in \{0, \dots, d-k-1\}$, then $\sigma_k(x) = \binom{n-k-1}{d-k}$.

Proof. Let first $m \neq k$. In order to reach a contradiction, assume that there exists $k \in \{0,1,\ldots,d-1\}$ and $m \in \{0,\ldots,d-k-1\}$ such that $\sigma_k(x) > 0$ and $\sigma_m(x) > 0$. By the definition of σ_j (Definition 10), we have that there exists a data-determined k-simplex U which contains x in its relative interior, and that there exists a data-determined m-simplex T which contains x in its relative interior. Let $y_1,\ldots,y_\ell$ be their common vertices, where $\ell \in \{0,\ldots,\min(k,m)\}$. By the definition of the relative interior it follows that x must belong to the set $(U \cap T) \setminus \operatorname{conv}(\{y_1,\ldots,y_\ell\})$. However, by Lemma 15 no such x exists.

Now, let $m = k$. As $\sigma_k(x)$ is nonzero, then $\sigma_k(x) \geq \binom{n-k-1}{d-k}$. Assume that $\sigma_k(x) > \binom{n-k-1}{d-k}$. Then there exists two different k-faces (k-simplices) U and T with ℓ common vertices $y_1,\ldots,y_\ell$, where $\ell \in \{0,\ldots,k\}$. The point x belongs to the relative interior of U and T, thus it follows that $x \in (U \cap T) \setminus \operatorname{conv}(\{y_1,\ldots,y_\ell\})$ which again contradicts Lemma 15.

□

To elucidate the meaning of the statements in Lemmas 15 and 16, we may look at the situation in $\mathbb{R}^4$.

Example 2.4. Let us have data points in $\mathbb{R}^4$ in general position. Consider a 1-simplex T_1 and a 2-simplex T_2 induced by some of these data points. In terms of Lemma 15 we have $k = 1$, $m = 2$, $d = 4$, $\ell \in \{0,1,2\}$, and

$$k + m - \ell = 1 + 2 - \ell \leq 3 = d - 1, \qquad \text{for all } \ell \in \{0,1,2\}.$$

Thereby, the intersection of T_1 and T_2 is either equal to the convex hull of their common vertices or an empty set. This further implies that $(\operatorname{relint} T_1) \cap (\operatorname{relint} T_2) = \emptyset$. Thus, for any position $x_1 \in \operatorname{relint} T_1$ we would have $\sigma_1(x_1) > 0$ and therefore $\sigma_2(x_1) = 0$.

Consider now two data-determined 2-simplices with no common vertices. Here we have $k = 2$, $m = 2$, $d = 4$, $\ell = 0$ and

$$k + m - \ell = 2 + 2 - 0 \not\leq 3 = d - 1.$$

Conditions in Lemma 15 are not satisfied, thus we cannot assume that the intersection of the relative interior of these two data-determined 2-simplices is an empty set. Indeed, there exists a configuration of 6 vertices in $\mathbb{R}^4$ which lie in general position such that the relative interior of the corresponding 2-simplices intersect. Take for instance 2-simplices S_1 and S_2, where

$$S_1 = S\left(\begin{pmatrix}1\\0\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\\0\end{pmatrix}, \begin{pmatrix}-1\\-1\\0\\0\end{pmatrix}\right), \qquad S_2 = S\left(\begin{pmatrix}0\\0\\1\\0\end{pmatrix}, \begin{pmatrix}0\\0\\0\\1\end{pmatrix}, \begin{pmatrix}0\\0\\-1\\-1\end{pmatrix}\right).$$

Then, $S_1 \cap S_2 = (0,0,0,0)^\top$, but also $(\operatorname{relint} S_1) \cap (\operatorname{relint} S_2) = (0,0,0,0)^\top$. Besides, all of $n = 6$ vertices of S_1 and S_2 lie in a general position. In terms of Lemma 16, for a dataset comprised of the vertices of S_1 and S_2 and a position $o = (0,0,0,0)^\top$ we have that $\sigma_2(o)$ is nonzero as o belongs to the relative interior of both 2-simplices S_1 and S_2. Lemma 16 could yield $\sigma_k(o) = \sigma_2(o) = \binom{6-2-1}{4-2} = \binom{n-k-1}{d-k}$ if its conditions were satisfied. However, in this case, those conditions are not met since $m = k$ but $m \notin \{0,1\} = \{0,\ldots,d-k-1\}$. And indeed we can write

$$\sigma_k(o) = \sigma_2(o) \geq \binom{6-2-1}{4-2} + \binom{6-2-1}{4-2} > \binom{6-2-1}{4-2} = \binom{n-k-1}{d-k},$$

where the first inequality follows from the fact that there are $\binom{6-2-1}{4-2}$ different simplices with three of its vertices from S_1, and similarly $\binom{6-2-1}{4-2}$ different simplices with three of its vertices from S_2 and all such simplices contribute to $\sigma_2(o)$. $\triangle$

As Theorem 17 is also rather technical, we will endeavor to illuminate the idea behind its formulation in advance. Consider again Example 2.3 for $d = 3$, where we had a dataset $\{O,V,D_1, D_2, D_3\}$. The only positions that will be of our interest this time are those lying in the line segment $\overline{O,V}$. We will add one and then two data points to the original dataset (see Figure 2.7) and recalculate the new depth of these positions (see Figure 2.8). The points are added so that the general position is maintained and so that the data points O and V remain as extreme points of the convex hull of all data points. To clarify, an extreme point of a convex set is a point that lies on the boundary of the set and cannot be expressed as a convex combination of any other points in the set. As this is just a motivation for the upcoming theorem, the details, such as the exact location of the added points, will not be discussed. Note that thanks to Lemma 16 and Lemma 14 we only need to calculate the depth of

(i) positions lying at the intersection of some $(d-1)$-simplex and $\overline{O,V}$;

(ii) one representative of the positions lying between the intersections from (i).

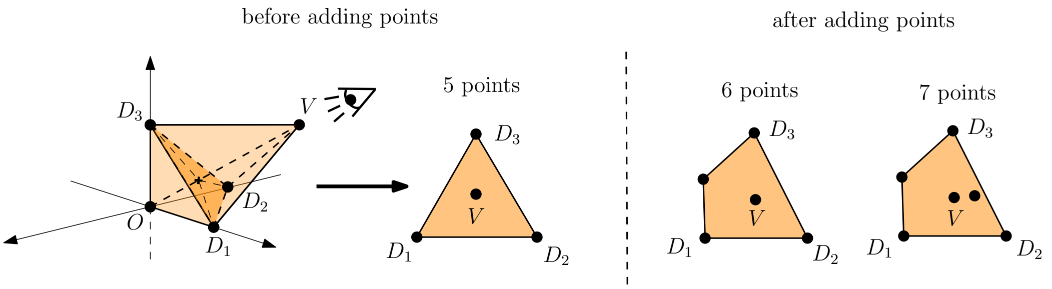


Figure 2.7: The original dataset with a different point of view, so that data points O and V overlap (left) and the new dataset after points "around" the line segment $\overline{O,V}$ have been added (right). Whether the new points are added within or outside the convex hull of the previous points is irrelevant to the observation.

To calculate the depth of positions of our concern, we use both SD_n^{closed} and SD_n^{avg}. However, only the results, not the actual calculation is provided in Figure 2.8. The sole purpose of including the depths calculated with SD_n^{closed} was to compare them to those obtained using SD_n^{avg}. Although SD_n^{closed} provided a more appropriate sample simplicial median for the dataset with 5 points, adding more points to the dataset seemingly (Figure 2.8) favored the use of SD_n^{avg}. In conclusion, while there may be situations where SD_n^{closed} produces better results, we conjecture that SD_n^{avg} would be generally more suitable for most applications.

The main purpose of this motivation, however, was to outline an interesting observation. The first thing to note is that none of these considered positions fulfills ***(C2)*** as they lie in $(d-2)$-simplex (line segment) $S(O,V)$. Therefore, Theorem 12 does not apply to them in any way. In spite of that, it appears that SD_n^{avg} of positions at the intersections can be obtained by averaging the depths

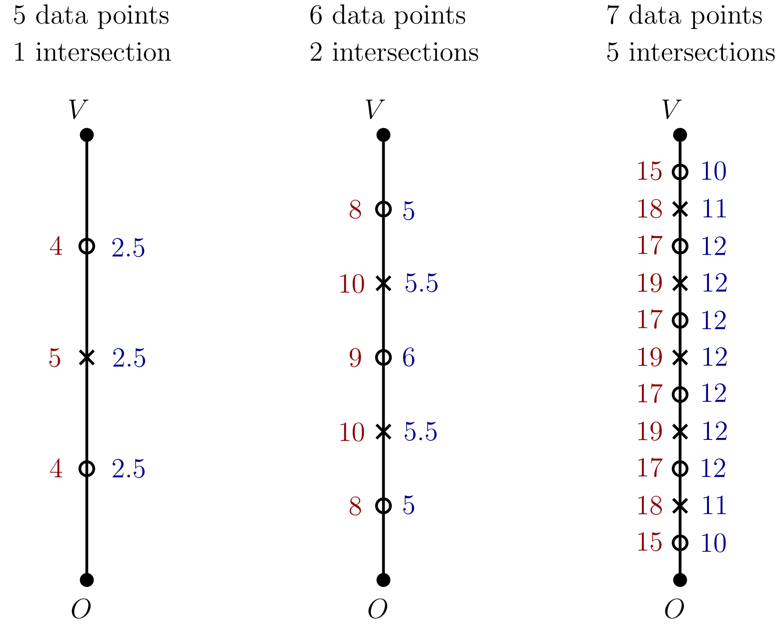


Figure 2.8: The depth of positions in the line segment $\overline{O,V}$ calculated using SD_n^{closed} (red) and with SD_n^{avg} (blue). The number of points in the dataset is displayed above each line segment. The cross symbols denote the intersections described in (i). Circles refer to the representatives from (ii).

of the closest representatives, one above and one below. The fact that this is not a coincidence is supported by the following theorem, which formalizes the generalization of this observation.

Theorem 17 (Addition to Theorem 12)**.** *Let $x_1,\dots,x_n \in \mathbb{R}^d$ be observed data points in general position. A hyperplane induced by some data-determined $(d-1)$-simplex D defines two* <u>*open*</u> *halfspaces H_1 and H_2. Let K be a data-determined k-simplex such that*

$$I_1 = (\operatorname{relint} K) \cap H_1 \neq \emptyset, \quad I_2 = (\operatorname{relint} K) \cap H_2 \neq \emptyset, \quad I = (\operatorname{relint} K) \cap (\operatorname{relint} D) \neq \emptyset,$$

where $k \in \{1,2,\dots,d-2\}$ (if such K exists). Additionally, let us assume the following

(A1) *For all $\eta \in I$ let $\sigma_{d-1}(\eta) = n-d$;*

(A2) *If $k \geq 2$, for all $\eta \in I$ let $\sigma_j(\eta) = 0$ for all $j \in \{d-k,\dots,d-2\}$.*

For $\varepsilon > 0$, let us denote $I^\varepsilon = \bigcup_{x\in I} B(x,\varepsilon)$, where $B(x,\varepsilon)$ is an open ball in $\mathbb{R}^d$ centered at x with radius ε. Then there exists $\varepsilon > 0$ small enough so that the depth of any position $\eta \in I$ is equal to the average of the depths of positions $y_1 \in I^\varepsilon \cap I_1$ and $y_2 \in I^\varepsilon \cap I_2$.

Proof. By assumption ***(A1)*** we have that $\sigma_{d-1}(\eta) = n-d$ which means that there is only one data-determined $(d-1)$-simplex that intersects I. There are only finitely many data-determined m-simplices $D_1,\dots,D_N$, $N \in \mathbb{N}$ (other than D) for $m = d-k,\dots,d-1$. Set

$$\varepsilon = \frac{1}{2} \cdot \min\{\operatorname{dist}(D_1,I),\dots,\operatorname{dist}(D_N,I)\},$$

where $\operatorname{dist}(A,B) = \inf\{\|x-y\| : x \in A, y \in B\}$ denotes the distance between sets A and B. The assumption ***(A1)*** cannot be satisfied if $\boldsymbol{\varepsilon}$ is zero, thus $\boldsymbol{\varepsilon} > 0$. For this $\boldsymbol{\varepsilon}$ we have $\sigma_{d-1}(y_1) = \sigma_{d-1}(y_2) = 0$, for y_1 and y_2 as in the statement of the theorem. By Lemma 14, Lemma 16 and assumption ***(A2)*** we have that $\sigma_k(\eta) = \sigma_k(y_1) = \sigma_k(y_2) = \binom{n-k-1}{d-k}$ and also that $\sigma_i(\eta) = \sigma_i(y_1) = \sigma_i(y_2) = 0$ for all $i \in \{0,1,2,\dots,d-2\} \setminus \{k\}$. Now we continue in a fashion similar to the proof of Theorem 12. Every simplex contributing to $\rho(\eta)$ also contributes to $\rho(y_1)$ and $\rho(y_2)$. Every simplex contributing to $\sigma_{d-1}(\eta)$ contains either y_1 or y_2 in its interior. Finally, every simplex which contribute to $\mathsf{SD}_n^{avg}(y_1)$ or $\mathsf{SD}_n^{avg}(y_2)$ must also contribute to $\mathsf{SD}_n^{avg}(\eta)$. Putting everything together and using the same notation as in the proof of Theorem 12 we get

$$\begin{aligned}\mathsf{SD}_n^{avg}(\eta) &= \frac{1}{2}(2\rho(\eta) + \sigma_{d-1}(\eta) + \sigma_k(\eta)) \\ &= \frac{1}{2}(\rho^{y_1}(\eta) + \rho^{y_2}(\eta) + \sigma_{d-1}^{y_1}(\eta) + \sigma_{d-1}^{y_2}(\eta) + \frac{1}{2}(\sigma_k(y_1) + \sigma_k(y_2))) \\ &= \frac{1}{2}(\mathsf{SD}_n^{avg}(y_1) + \mathsf{SD}_n^{avg}(y_2)),\end{aligned}$$

which completes the proof.

□

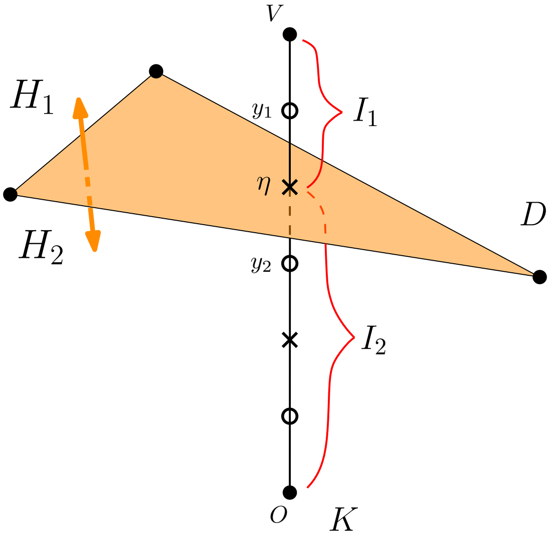


Figure 2.9: Theorem 17 applied on our motivational example in Figure 2.8 (6 data points) with $d = 3$: The $(d-1)$-simplex D (orange triangle) intersects the 1-simplex $K = S(O,V)$ in η. Thus, in this case $I = \{\eta\}$ is a singleton. Open halfspaces H_1 and H_2 are above and below D, respectively (orange arrows). Consequently, I_1 and I_2 are also above and below, respectively (red curly brackets). Note that both conditions ***(A1)*** and ***(A2)*** are satisfied. Condition ***(A2)*** is trivially true since $k = 1$. Thus, Theorem 17 can be applied, which gives a general version of the observation from Figure 2.8.

The combination of Theorems 12 and 17 enables us to derive a weakened generalized version of our Theorem 13, or conversely, an enhanced accurate version of the incorrect Proposition 10.

Theorem 18 (Proposition 10, $d \geq 2$)**.** *Let $x_1, \ldots, x_n \in \mathbb{R}^d$ be observed data points in general position. Let H denote the convex hull formed by these points. Then the following statements hold true:*

(i) In order to find the maximum averaged sample simplicial depth, it is not necessary to calculate the depth of these types of positions: θ in the sense of Theorem 12, and η in the sense of Theorem 17.

(ii) If the observed data points are not in a convex position[8], then the maximum averaged sample simplicial depth is not attained at the boundary of H.

Proof. (i) By Theorems 12 and 17, whenever the depth of a sample simplicial median is attained at any of the positions of these types (θ, η), then it must be also attained at some positions which do not fall into these categories.

(ii) The depth of any position from the boundary of H is always less then the depth of any boundary data point. By the assumption that the observed data points are not in a convex position, there exist an index i such that x_i is contained in the interior of H. By the Carathéodory theorem [18, Theorem 2.2.4], at least one simplex with boundary data points as vertices contains x_i in its interior. We can write

$$\mathsf{SD}_n^{avg}(x_i) \geq 1 + \frac{1}{2}\binom{n-1}{d} > \frac{1}{2}\binom{n-1}{d} = \mathsf{SD}_n^{avg}(x_j),$$

where x_j is an arbitrary boundary data point.

□

Note. In contrast to Theorem 13 (ii), for general d, we only managed to prove a similar assertion for data points that are not in convex position. Even though the problem for data points in convex position may seem intuitively trivial, we were not able to come up with any formal proof.

One of the ideas to approach that problem was to apply the so called first selection lemma from discrete geometry [13, Chapter 9] which in our case can be formulated as follows.

Lemma 19 (First selection lemma)**.** *Consider a dataset of n points in $\mathbb{R}^d$. Then there exists a point $p \in \mathbb{R}^d$ contained in at least $c_d \cdot \binom{n}{d+1}$ data-determined simplices, where $c_d > 0$ is constant depending only on the dimension d.*

It should be pointed out that the best possible value for c_d is not known, except for $d = 2$. For $d \geq 3$ only bounds are available. Let us look at the situation in $d = 3$. It was proved in [1] that for a set M of n points in $\mathbb{R}^3$ there exists a point $p \in \mathbb{R}^3$ contained in at least $0.00227 \cdot n^4$ simplices spanned by M. In our case, however, we need to differentiate between simplices which contain p in their interior and those which contain p in their boundary. To do so, we can use a lemma from [13, Lemma 9.1.2] which states that for set N of n points in general position, no point in $\mathbb{R}^d$ is contained in more than dn^{d-1} hyperplanes

[8] A set of points is said to lie in a convex position if none of the points can be represented as a convex combinations of the others.

induced by N. The lower bound derived from these assertions would lead to the following inequality

$$0.00227 \cdot n^4 - 3 \cdot n^2 + \frac{3}{2} \cdot n^2 > \frac{1}{2}\binom{n-1}{3},$$

which holds true for $n > 46$. Consequently, for $d = 3$ and $n > 46$, we can guarantee that the maximum averaged sample simplicial depth is not attained at the boundary of H from Theorem 18. The problem of determining whether there exists a configuration of n data points in $\mathbb{R}^3$ with $n \leq 46$, where the maximum averaged sample simplicial depth is highest at the boundary, remains unresolved.

During the search for such configuration, we explored the following idea.

Example 2.5. Consider general dimension $d \geq 3$ and a simplex S. Let us have $(d+1)$ groups of k data points, where each group is situated in the neighborhood of one of the vertices of S, so that the data points lie in convex position (for $d = 3$ see Figure 2.10). Together we have $n = k \cdot (d+1)$ data points.

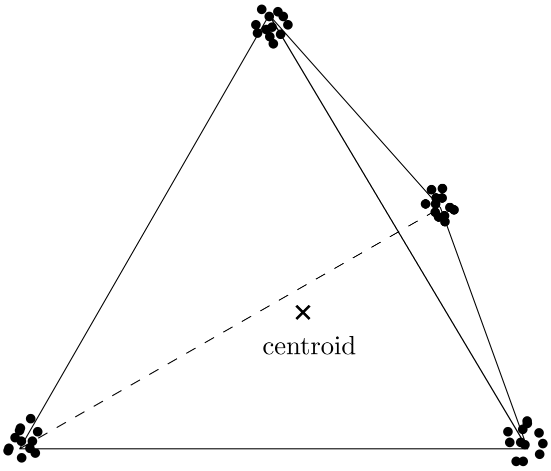


Figure 2.10: Situation in Example 2.5 for $d = 3$ and $k = 12$.

In this specific configuration we can easily determine the depth SD_n^{avg} of

- any data point D: $\mathsf{SD}_n^{avg}(D) = \frac{1}{2}\binom{n-1}{d} = \frac{1}{2}\binom{(d+1)k-1}{d}$;
- any position M in the "middle" cell: $\mathsf{SD}_n^{avg}(M) = k^{d+1}$, where by the middle cell we understand the cell which contains the centroid of S.

The latter follows from the observation that M is contained in exactly those data-determined simplices whose all vertices belong to different k-element groups. Notice, that $\mathsf{SD}_n^{avg}(D)$ is $\mathcal{O}(n^d)$, while $\mathsf{SD}_n^{avg}(M)$ is $\mathcal{O}(n^{d+1})$. Hence, for n large enough we have $\mathsf{SD}_n^{avg}(D) < \mathsf{SD}_n^{avg}(M)$, which corresponds to the first selection lemma, as we would expect. However, for n small enough we have $\mathsf{SD}_n^{avg}(D) > \mathsf{SD}_n^{avg}(M)$. For instance, in dimension $d = 3$, the inequality

$$\mathsf{SD}_n^{avg}(D) = \frac{1}{2}\binom{4k-1}{3} > k^4 = \mathsf{SD}_n^{avg}(M)$$

holds for $k = 2$ or $k = 3$. Meanwhile, in dimension $d = 25$, the inequality

$$\mathsf{SD}_n^{avg}(D) = \frac{1}{2}\binom{26k-1}{25} > k^{26} = \mathsf{SD}_n^{avg}(M)$$

holds for every $k \leq 7632345541$. Due to these calculations we had a hypothesis which goes as follows: For each $d \geq 3$ there exists $n > d+1$ such that for less than n observed data points in convex position, there exists a configuration such that the maximum averaged sample simplicial depth is attained in a data point.

It is important to note that the depth of position M being smaller than the depth of data point D does not necessarily mean that the maximum averaged sample simplicial depth is attained in D. Therefore, to test our hypothesis we have to calculate the depth of other positions as well. In order to do so, we use the "easiest to compute" situation. That is, a dataset of 8 points in $\mathbb{R}^3$ located as described in the beginning of this example (Figure 2.10 with $k = 2$ instead of $k = 12$). At first, in order to compute SD_8^{avg} in each (most) of the cells, we randomly generated 10000 positions inside the convex hull of all 8 data points. The maximum SD_8^{avg} of these generated positions was 17 which is less than $\mathsf{SD}_8^{avg}(D) = 17.5$. Thus, it is highly unlikely there is a cell with greater depth than the depth of a data point. By Theorem 12, the same remark can be made for positions in the relative interior of some data-determined 2-simplex. The only positions left, whose depth we need to evaluate, are those lying in data-determined 1-simplices. Due to the general position of data points, no two data-determined 1-simplices intersect. Besides, thanks to Theorem 17 we do not need to calculate the depth of positions which have nonzero both σ_2 and σ_1. This enables us to bypass the need to search for intersections among data-determined 1-simplices and 2-simplices. By randomly generating positions within data-determined 1-simplices we found that the maximum SD_8^{avg} value was 21.5, exceeding 17.5. As a result, we ended up not being able to support our hypothesis in this particular case, but our general hypothesis remains open. △

3 A paradox

While writing this thesis we stumbled upon an interesting problem which goes by the name Sylvester's four-point problem [17, 10]. It can be stated as follows: *"What is the probability that four randomly chosen points in $\mathbb{R}^2$ create a convex quadrilateral?"* Due to the inaccurate phrasing of the original question, many inconsistent results were published throughout the years. For an overview see [15]. Clearly, the question depends on the considered probability distribution, or the shape of the subset of $\mathbb{R}^2$ from which the points are randomly sampled if the distribution is uniform. Most of the results focus on points sampled independently from continuous uniform distribution defined on some convex compact subset of $\mathbb{R}^2$ with non-empty interior. Not so long ago, results concerning a Gaussian distribution were published as well [3]. For some of the most recent results, regarding a discrete uniform distribution defined on an $n \times n$ net in $\mathbb{R}^2$ we refer to [10].

For the purposes of this chapter, only continuous uniform distributions defined on a convex compact subset K of $\mathbb{R}^2$ are of our concern and we will denote them by $P_K \in \mathcal{P}(\mathbb{R}^2)$. Let us further denote the following complementary events

$CQ_K =$ [random sample X_1, X_2, X_3, X_4 from P_K created a convex quadrilateral];
$NQ_K = \Omega \setminus CQ_K$.

In order to find the probability $\mathsf{P}(CQ_K)$, it is sufficient to determine the expected value of the area spanned by three random points from P_K. We can see that from

$$\begin{aligned}\mathsf{P}(CQ_K) &= 1 - \mathsf{P}(NQ_K)\\ &= 1 - \mathsf{P}(X_1 \in S(X_2,X_3,X_4)) - \mathsf{P}(X_2 \in S(X_1,X_3,X_4))\\ &\quad - \mathsf{P}(X_3 \in S(X_1,X_2,X_4)) - \mathsf{P}(X_4 \in S(X_1,X_2,X_3))\\ &= 1 - 4\mathsf{P}(X_4 \in S(X_1,X_2,X_3))\\ &= 1 - 4 \cdot \frac{\mathsf{E}_{X_1,X_2,X_3}[\lambda^2(S(X_1,X_2,X_3))]}{\lambda^2(K)},\end{aligned}$$

where $\lambda^2(\cdot)$ denotes Lebesgue measure. In the last equality we used a geometric interpretation of the continuous uniform distribution P_K. This allows us to establish a connection with the simplicial depth from Definition 6

$$\begin{aligned}\mathsf{E}_{X_4}[SD(X_4; P_K)] &= \mathsf{E}_{X_4}[P(X_4 \in S(X_1, X_2, X_3))]\\ &= \mathsf{E}_{X_4}\left[\frac{\mathsf{E}_{X_1,X_2,X_3}[\lambda^2(S(X_1,X_2,X_3))]}{\lambda^2(K)}\right]\\ &= \frac{\mathsf{E}_{X_1,X_2,X_3}[\lambda^2(S(X_1,X_2,X_3))]}{\lambda^2(K)}.\end{aligned}$$

To put this in words, the expected value of simplicial depth $SD(X; P_K)$ is the same as the expected value of the area spanned by three random points from P_K, all that divided by the area of K.

Interestingly enough, Blaschke in [2] managed to prove that for all convex compact $K \subseteq \mathbb{R}^2$

$$\mathsf{E}_{X_4}[SD(X_4; P_{ellipse})] = \frac{35}{48\pi^2} \leq \mathsf{E}_{X_4}[SD(X_4; P_K)] \leq \frac{1}{12} = \mathsf{E}_{X_4}[SD(X_4; P_{triangle})].$$

Here $P_{ellipse}$ stands for the uniform distribution defined on (any full-dimensional) ellipse in $\mathbb{R}^2$, and $P_{triangle}$ analogously for a triangle. As we can see, the mean value of the simplicial depth is maximal for $P_{triangle}$, while minimal for $P_{ellipse}$. On the other hand, if we now compare the maximum simplicial depth of $P_{triangle}$ and $P_{ellipse}$ instead, we conjecture that

$$\max_{x \in \, triangle} SD(x; P_{triangle}) = \frac{3}{729}(18 + 40 \cdot \log(2) + 5 \cdot \log(16)) \approx 0.245222, \quad (3.1)$$

which would result in a converse inequality as it is less than

$$\max_{x \in \, ellipse} SD(x; P_{ellipse}) = \frac{1}{4}. \quad (3.2)$$

Equation (3.2) follows from Theorem 6 as $P_{ellipse}$ is C-symmetric about the center of the considered ellipse. The distribution of $P_{triangle}$, however, is not even H-symmetric about any point. Therefore, Theorem 6 cannot be used.

Naturally, one would expect the simplicial median of the distribution $P_{triangle}$ to be located at the centroid of the considered triangle. First of all, let us mention that in order to show that simplicial median is attained in the centroid of a triangle, it is immaterial what non-degenerate triangle we consider. This claim follows from the affine invariance in Theorem 3. Therefore, we proceed by considering an equilateral triangle T with vertices $A = (-\frac{1}{2}, -\frac{1}{2\sqrt{3}})^\top$, $B = (\frac{1}{2}, -\frac{1}{2\sqrt{3}})^\top$, $C = (0, \frac{1}{\sqrt{3}})^\top$. The length of the side of T is 1 and the centroid M of T is located at $(0,0)^\top$. By the statement (1.5) in the corollary of our treatment of consistency of the simplicial depth, we know that $SD_n^{closed}(x; P_n(\omega))$ converges to $SD(x; P_T)$ almost surely. Thus, to support our hypothesis of $SD(M; P_T) = \max_{x \in T} SD(x; P_T)$, we randomly generated 10000 points uniformly inside T and calculated $SD_{10000}^{closed}(\cdot; \hat{P}_{10000})$ for all generated points, where $\hat{P}_{10000}$ is the empirical measure induced by the generated points. The maximum sample simplicial depth was attained at the point $p \approx (0.000, 0.005)^\top$ with $SD_{10000}^{closed}(p; \hat{P}_{10000}) \approx 0.2454$. To increase our level of certainty even further, we randomly generated additional 2500 points uniformly in a circular disk centered at M with radius 0.07. After computing the corresponding depth of these 2500 points with respect to $\hat{P}_{10000}$, we again obtained nearly identical results as before. For a visualization of this simulation exercise, see Figure 3.1. Unfortunately, as we were not able to rigorously prove that the centroid of the triangle is the point where the simplicial median is attained, we were bound to use the word "conjecture."

Before we proceed to the value of the depth presented in the equation (3.1), it is worth noting that computing the simplicial depth SD with respect to a probability measure $P \in \mathcal{P}(\mathbb{R}^d)$ is a challenging task. As far as we are aware, there are only three types of probability distributions P whose simplicial depth has been exactly determined in the literature:

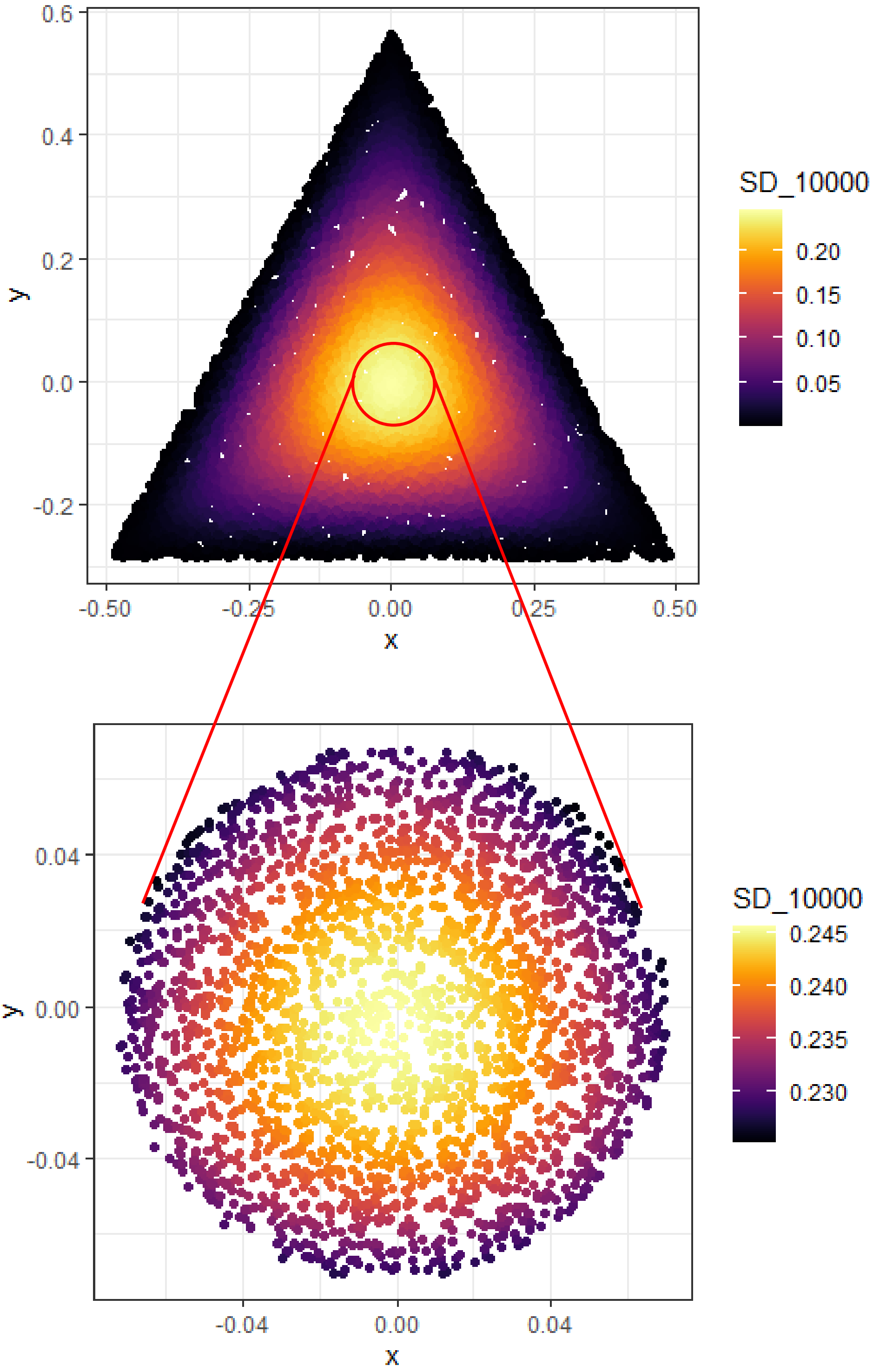


Figure 3.1: The top figure displays 10000 randomly generated points. The depth of these points is represented with colors, according to the legend to its right. The bottom figure displays another 2500 randomly generated points, where again, their depth is represented with colors corresponding to the legend to its right. Clearly, the simulation suggests that the simplicial median is indeed attained at M.

- One-dimensional measures $P \in \mathcal{P}(\mathbb{R})$;
- Finitely supported measures $P \in \mathcal{P}(\mathbb{R}^d)$;
- Several specific measures $P \in \mathcal{P}(\mathbb{R}^d)$, whose support is a one-dimensional subset of $\mathbb{R}^d$.

For more detailed discussion we refer to [14, Chapter 3]. The complexity of determining the probability in (1.1) (for $d \geq 2$) stems from the involved integrals being too complicated to be solved directly. However, during our research, we developed a method in $\mathbb{R}^2$ to reduce this problem to evaluating a single integral of a function with one variable. Although this simplification is rather intriguing, we will not include it due to time and space constraints of this work. Hopefully, we will be able to provide a detailed explanation elsewhere.

Thanks to this method, we were able to determine the exact simplicial depth of several points lying on the median (in the sense of a median line of a triangle) of T with respect to P_T. The results are plotted in Figure 3.2. Besides, the results support our conjecture of centroid M being the point with the greatest depth.

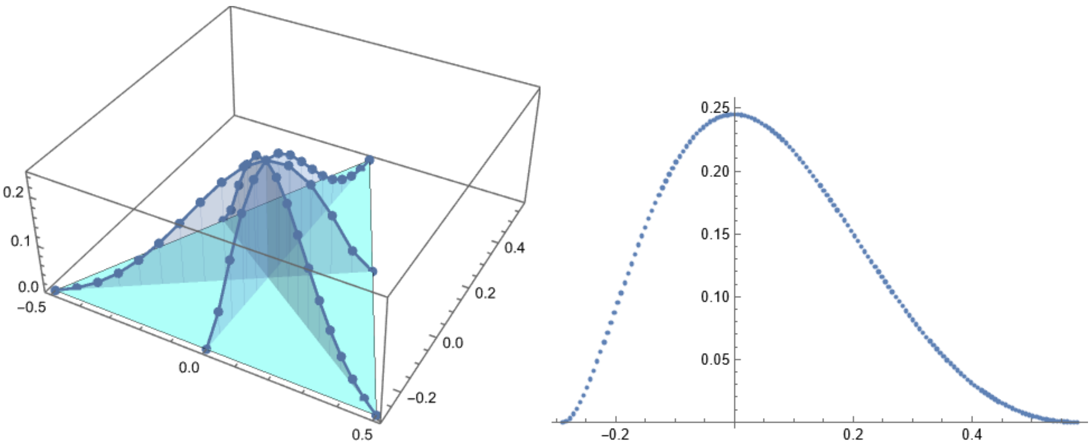


Figure 3.2: (left): triangle T depicted in a 3-dimensional space (cyan). The z coordinate of 46 dark blue points represents the simplicial depth of the corresponding points lying on the medians of T. The points were connected for visualization purposes. (right): Here, the value x on the horizontal axis determines point $(0,x)^\top$, and the value y on the vertical axis represents a simplicial depth with respect to P_T.

Hence, according to our calculations, if we were to assume that M is the simplicial median, we could write

$$\max_{x \in T} SD(x; P_T) = SD(M; P_T) = \frac{3}{729}(18 + 40 \cdot \log(2) + 5 \cdot \log(16)).$$

Finally, as the inequalities

$$\begin{gathered} \mathsf{E}_{X_4}[SD(X_4; P_{ellipse})] < \mathsf{E}_{X_4}[SD(X_4; P_{triangle})]; \\ \max_{x \in ellipse} SD(x; P_{ellipse}) > \max_{x \in triangle} SD(x; P_{triangle}), \end{gathered}$$

seems counter-intuitive at the first glance, we decided to label them as a paradox.

Conclusion

In this work, we introduced and thoroughly examined the concept of simplicial depth, providing detailed proofs of its four main properties. Our proofs were accompanied by illustrations, making them more accessible.

In the second part of the thesis we explored various definitions of sample simplicial depth and focused primarily on the averaged sample simplicial depth. We corrected and reformulated several original propositions while extending parts of the related theory further.

Our efforts led us to an intriguing problem: *Is there a configuration of data points in dimension $d \geq 3$ in convex position such that the maximum averaged sample simplicial depth is attained in a data point?* Although this problem may appear simple at first, we were unable to uncover a solution. Furthermore, we proposed that this problem is closely tied to the unsolved first selection lemma for $d \geq 3$, or a version of it, where we only consider points in convex position.

Quite recently, we discovered a potential original link between the simplicial depth and Sylvester's four-point problem. That novel aspect of simplicial depth could be promising for determining the exact simplicial depth for absolutely continuous distributions in $\mathbb{R}^2$. However, due to time limitations, we were unable to fully explore the practical implications of this connection, and only presented an interesting paradox that we encountered.

In any case, the study of simplicial depth still holds many unanswered questions, and we plan to pursue further investigations in this area in the near future.

Bibliography

[1] Abdul Basit, Nabil H. Mustafa, Saurabh Ray, and Sarfraz Raza. Improving the first selection lemma in $\mathbb{R}^3$. In *Computational geometry (SCG'10)*, pages 354–357. ACM, New York, 2010.

[2] W. Blaschke. Über affine Geometrie XI: Lösung des "Vierpunktproblems" von Sylvester aus der Theorie der geometrischen Wahrscheinlichkeiten. *Leipziger Berichte*, 69:436–453, 1917.

[3] Christian Blatter. Four shots for a convex quadrilateral. *Amer. Math. Monthly*, 115(9):837–843, 2008.

[4] Michael A. Burr, Eynat Rafalin, and Diane L. Souvaine. Simplicial depth: An improved definition, analysis, and efficiency for the finite sample case. In *Canadian Conference on Computational Geometry*, pages 136–139, 2003.

[5] Michael A. Burr, Eynat Rafalin, and Diane L. Souvaine. Simplicial depth: An improved definition, analysis, and efficiency for the finite sample case. In *Data depth: Robust multivariate analysis, computational geometry and applications*, volume 72 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pages 195–209. Amer. Math. Soc., Providence, RI, 2006.

[6] Zhiqiang Chen. Bounds for the breakdown point of the simplicial median. *J. Multivariate Anal.*, 55(1):1–13, 1995.

[7] Víctor H. de la Peña and Evarist Giné. *Decoupling.* Probability and its Applications. Springer-Verlag, New York, 1999.

[8] R. M. Dudley. *Real analysis and probability*, volume 74 of *Cambridge Studies in Advanced Mathematics.* Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original.

[9] Lutz Dümbgen. Limit theorems for the simplicial depth. *Statist. Probab. Lett.*, 14(2):119–128, 1992.

[10] Eliška Hálová. Problém čtyř bodů, 2023. Bc. thesis, Matematicko-fyzikální fakulta Univerzity Karlovy.

[11] Regina Y. Liu. On a notion of simplicial depth. *Proc. Nat. Acad. Sci. U.S.A.*, 85(6):1732–1734, 1988.

[12] Regina Y. Liu. On a notion of data depth based on random simplices. *Ann. Statist.*, 18(1):405–414, 1990.

[13] Jiří Matoušek. *Lectures on discrete geometry*, volume 212 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2002.

[14] Stanislav Nagy. Simplicial depth and its median: Selected properties and limitations. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, to appear, 2023.

[15] Richard E. Pfiefer. The historical development of J. J. Sylvester's four point problem. *Math. Mag.*, 62(5):309–317, 1989.

[16] Peter J. Rousseeuw and Anja Struyf. Characterizing angular symmetry and regression symmetry. *J. Statist. Plann. Inference*, 122(1-2):161–173, 2004.

[17] J.J Sylvester. *The Educational Times, Question 1491, London*, 1864.

[18] Roger Webster. *Convexity*. Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, 1994.

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

[20] Yijun Zuo and Robert Serfling. On the performance of some robust nonparametric location measures relative to a general notion of multivariate symmetry. *J. Statist. Plann. Inference*, 84(1-2):55–79, 2000.

[21] Adam Říha. Symetrie náhodných vektorů, 2021. Mgr. thesis, Matematicko-fyzikální fakulta Univerzity Karlovy.