Extending the Centerpoint Theorem to Multiple Points

Alexander Pilz[1]

Institute of Software Technology, Graz University of Technology, Austria
apilz@ist.tugraz.at
https://orcid.org/0000-0002-6059-1821

Patrick Schnider

Department of Computer Science, ETH Zurich, Switzerland
patrick.schnider@inf.ethz.ch

Abstract

The centerpoint theorem is a well-known and widely used result in discrete geometry. It states that for any point set P of n points in $\mathbb{R}^d$, there is a point c, not necessarily from P, such that each halfspace containing c contains at least $\frac{n}{d+1}$ points of P. Such a point c is called a centerpoint, and it can be viewed as a generalization of a median to higher dimensions. In other words, a centerpoint can be interpreted as a good representative for the point set P. But what if we allow more than one representative? For example in one-dimensional data sets, often certain quantiles are chosen as representatives instead of the median.

We present a possible extension of the concept of quantiles to higher dimensions. The idea is to find a set Q of (few) points such that every halfspace that contains one point of Q contains a large fraction of the points of P and every halfspace that contains more of Q contains an even larger fraction of P. This setting is comparable to the well-studied concepts of weak ε-nets and weak ε-approximations, where it is stronger than the former but weaker than the latter. We show that for any point set of size n in $\mathbb{R}^d$ and for any positive $\alpha_1, \ldots, \alpha_k$ where $\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_k$ and for every i, j with $i + j \leq k + 1$ we have that $(d-1)\alpha_k + \alpha_i + \alpha_j \leq 1$, we can find Q of size k such that each halfspace containing j points of Q contains least $\alpha_j n$ points of P. For two-dimensional point sets we further show that for every α and β with $\alpha \leq \beta$ and $\alpha + \beta \leq \frac{2}{3}$ we can find Q with $|Q| = 3$ such that each halfplane containing one point of Q contains at least αn of the points of P and each halfplane containing all of Q contains at least βn points of P. All these results generalize to the setting where P is any mass distribution. For the case where P is a point set in $\mathbb{R}^2$ and $|Q| = 2$, we provide algorithms to find such points in time $O(n \log^3 n)$.

2012 ACM Subject Classification Theory of computation → Design and analysis of algorithms

Keywords and phrases centerpoint, point sets, Tukey depth

Digital Object Identifier 10.4230/LIPIcs.ISAAC.2018.53

Related Version A full version of the paper is available at `https://arxiv.org/abs/1810.10231`.

Acknowledgements We thank Emo Welzl for initiating discussions on this topic, as well as anonymous reviewers for helpful comments.

[1] Supported by a Schrödinger fellowship of the Austrian Science Fund (FWF): J-3847-N35.

29th International Symposium on Algorithms and Computation (ISAAC 2018).
Editors: Wen-Lian Hsu, Der-Tsai Lee, and Chung-Shou Liao; Article No. 53; pp. 53:1–53:13
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Medians and quantiles are ubiquitous in the statistical analysis and visualization of data. These notions allow for quantifying how deep some point lies within a one-dimensional data set by measuring how many elements of the data set appear before the point and how many appear after it. In comparison to the mean, medians and quantiles have the advantage that they only depend on the order of the data points, and not their exact positions, making them robust against outliers. However, in many applications, data sets are multidimensional, and there is no clear order of the data set. For this reason, various generalizations of medians to higher dimensions have been introduced and studied. Many of these generalized medians rely on a notion of depth of a query point within a data set, a median then being a query point with the highest depth among all possible query points. Several such depth measures have been introduced over time, most famously Tukey depth [18] (also called halfspace depth), simplicial depth, or convex hull peeling depth (see, e.g., [1]). All of these depth measures lead to generalized medians that are invariant under affine transformations. As for quantiles, only a few generalizations have been introduced (see, e.g., [6]). We propose such a generalization by extending a depth measure to sets with a fixed number of query points and defining a quantile as a set with maximal depth. The depth measure we extend is Tukey depth: the *Tukey depth* of a point q with respect to a point set $P \subset \mathbb{R}^d$ is the minimal number of points of P in any closed halfspace containing q. More formally, if H denotes the set of closed halfspaces, then the Tukey depth $\mathsf{td}_P(q)$ of q with respect to P is

$$\mathsf{td}_P(q) = \min_{q \in h \in H} \{|h \cap P|\} \ .$$

Similarly, the Tukey depth can also be defined for a mass distribution μ:

$$\mathsf{td}_\mu(q) = \min_{q \in h \in H} \{\mu(h)\} \ .$$

Here, a *mass distribution* μ on $\mathbb{R}^d$ is a measure on $\mathbb{R}^d$ such that all open subsets of $\mathbb{R}^d$ are measurable, $0 < \mu(\mathbb{R}^d) < \infty$ and $\mu(S) = 0$ for every lower-dimensional subset S of $\mathbb{R}^d$.

The centerpoint theorem states that there is always a point of high depth, i.e., a point q such that for every closed halfspace h containing q we have $|h \cap P| \geq \frac{|P|}{d+1}$ (or $\mu(h) \geq \frac{\mu(\mathbb{R}^d)}{d+1}$ for masses). Note that, for $d = 1$, such a centerpoint is a median: a median has the property that every halfline containing it contains at least half of the underlying data set. Quantiles can be interpreted similarly: the $\frac{1}{3}$-quantile and the $\frac{2}{3}$-quantile form a set of two points such that every halfline that contains one of them contains at least $\frac{1}{3}$ of the data set. Furthermore, a halfline containing both of the points contains at least $\frac{2}{3}$ of the underlying data set. In particular, halflines containing more points contain more of the data set. This idea leads to the following generalization of Tukey depth for a set Q of multiple points:

$$\mathsf{gtd}_P(Q) := \min_{h \in H\colon Q \cap h \neq \emptyset} \left\{ \frac{|h \cap P|}{|h \cap Q|} \right\} \ .$$

Again, we can generalize this to mass distributions:

$$\mathsf{gtd}_\mu(Q) := \min_{h \in H\colon Q \cap h \neq \emptyset} \left\{ \frac{\mu(h)}{|h \cap Q|} \right\} \ .$$

We prove that there is always a set Q of k points that has generalized Tukey depth $\frac{1}{kd+1}$. In fact, we prove the following, more general statement:

▶ **Theorem 1.** *Let μ be a mass distribution in $\mathbb{R}^d$ with $\mu(\mathbb{R}^d) = 1$. Let $\alpha_1, \ldots, \alpha_k$ be non-negative real numbers such that $\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_k$ and for every i, j with $i + j \leq k + 1$ we have that $(d-1)\alpha_k + \alpha_i + \alpha_j \leq 1$. Then there are k points $p_1, \ldots, p_k$ in $\mathbb{R}^d$ such that for each closed halfspace h containing j of the points $p_1, \ldots, p_k$ we have $\mu(h) \geq \alpha_j$.*

Note that, for $d = 1$ and $k = 2$, the points p_1 and p_2 correspond to the α_1-quantile and the $(1-\alpha_1)$-quantile; for $\alpha_j = \frac{j}{kd+1}$ we get our bound on the generalized Tukey depth, and for $\alpha_1 = \ldots = \alpha_k$, the result implies the centerpoint theorem.

Our second result is motivated by interpreting the $\frac{1}{3}$-quantile and the $\frac{2}{3}$-quantile not as two points, but as a one-dimensional simplex. We then have that every halfline that contains a part of the simplex contains at least $\frac{1}{3}$ of the underlying data set and every halfline that contains the whole simplex contains at least $\frac{2}{3}$ of the underlying data set. Also for this interpretation we give a generalization to two dimensions:

▶ **Theorem 2.** *Let μ be a mass distribution in $\mathbb{R}^2$ with $\mu(\mathbb{R}^2) = 1$. Let α and β be real numbers such that $0 < \alpha \leq \beta$ and $\alpha + \beta = \frac{2}{3}$. Then there is a triangle Δ in $\mathbb{R}^2$ such that*

(1) *for each closed halfplane h containing one of the vertices of Δ we have $\mu(h) \geq \alpha$ and*

(2) *for each closed halfplane h fully containing Δ we have $\mu(h) \geq \beta$.*

Note that this again generalizes centerpoints for $\alpha = \beta$. However, this result does not give bounds on the generalized Tukey depth of these sets, as, e.g., a halfspace containing two points may still only contain an α-fraction of the mass.

Finally, we give algorithms to compute two points satisfying the two-dimensional case of Theorem 1 and three points satisfying Theorem 2 in time $O(n \log^3 n)$.

Related work. Another way to view our setting is the following: given a multidimensional data set, we want to find a fixed number of representatives. The idea of small point sets representing a larger point set has been studied in many different settings. One of the most famous of those is the concept of ε-nets, introduced by Haussler and Welzl [7]. For a range space (X, R), consisting of a set X and a set R of subsets of X, an ε-net on $P \subset X$ is a subset N of P with the property that every $r \in R$ with $|r \cap P| \geq \varepsilon|P|$ intersects N. In our setting, where we consider halfspaces, we would choose $X = \mathbb{R}^d$ and R as the set of all halfspaces. It is known that for this range space, for any point set P there exists an ε-net of size $O(\frac{d}{\varepsilon} \log \frac{d}{\varepsilon})$. In particular, this bound does not depend on the size of P. Note that we require the ε-net to be a subset of P. If this condition is dropped, we arrive at the concept of *weak* ε-nets. The fact that the points for the weak ε-net can be chosen anywhere in $\mathbb{R}^d$ allows for very small weak ε-nets for many range spaces. There has been some work on weak ε-nets of small size. For halfplanes in $\mathbb{R}^2$ for example, Aronov et al. [3] have shown that there is always a weak $\frac{1}{2}$-net of two points. These two points both lie outside of the convex hull of P. They also consider many other range spaces, such as convex sets, disks and rectangles. Similarly, Babazadeh and Zarrabi-Zadeh [4] construct weak $\frac{1}{2}$-nets of size 3 for halfspaces in $\mathbb{R}^3$. For two-dimensional convex sets, Mustafa and Ray [15] have shown that there is always a weak $\frac{4}{7}$-net of two points; Shabbir [17] shows how to find two such points in $O(n \log^4 n)$ time.

Another related concept is the concept of ε-approximations: For a range space (X, R) an ε-approximation on $P \subset X$ is a subset N of P with the property that for every $r \in R$ we have $\left| \frac{|r \cap P|}{|P|} - \frac{|r \cap N|}{|N|} \right| \leq \varepsilon$. Again, the restriction that N has to be a subset of P can be dropped to get the notion of weak ε-approximations. Just as for ε-nets, there has been a lot of work on ε-approximations and weak ε-approximations, see [14] for a recent survey. In particular it was shown that for halfspaces in $\mathbb{R}^d$, there always is an ε-approximation of size $O(\frac{1}{\varepsilon^{2-2/(d+1)}})$ [12, 13].

While our setting can be considered to be related to weak ε-nets and weak ε-approximations for range spaces determined by halfspaces, the differences are significant. As we will discuss here, a halfspace missing all the points of Q may still contain up to half of the points of the initial set, and thus Q qualifies neither as a good weak ε-approximation nor ε-net.

Note that for $|Q| = 2$, the condition of Theorem 1 that any halfspace containing all of the points of Q contains at least $\alpha_2 n$ points of P is equivalent to the statement that every halfspace containing more than $(1 - \alpha_2)n$ of the points of P contains at least one point of Q. So, Q is a weak $(1 - \alpha_2)$-net of P. Furthermore, the condition that any halfspace containing one of the points of Q contains at least $\alpha_1 n$ points of P translates to the statement that every halfspace containing more than $(1 - \alpha_1)n$ of the points of P must contain all of Q. Thus, Q is not only a weak $(1 - \alpha_2)$-net of P but also has the additional property that all points of Q are somewhat deep within P. (For two points in the plane, this comes at the cost of having ε larger than $\frac{1}{2}$.) On the other hand, while we require halfspaces containing all points of Q to also contain many points of P, we also allow halfspaces containing only one point of Q to contain many points of P. This separates our concept from weak ε-approximations. Note that when dealing with halfspaces and ε-nets of size 2, the weak $\frac{1}{2}$-net of Aronov et al. [3] is actually also a weak $\frac{1}{2}$-approximation. Similarly, Theorem 1 gives us a weak $(1 - \alpha_2)$-approximation of P, with the optimal value being reached when α_1 tends to zero (which actually corresponds to the result in [3]). Adding more points to Q does not give us a better approximation. For $d = 2$, requiring that for $i + j \leq k + 1$ we have $(d - 1)\alpha_i + \alpha_j + \alpha_k < 1$ implies $\alpha_1 + 2\alpha_k < 1$, so a halfspace containing no points of Q may contain half of the points of P; we therefore cannot get anything better than a weak $\frac{1}{2}$-approximation. Also, we do not get anything better than a weak $\frac{1}{2}$-net.

In fact, our setting is very much related to the concept of one-sided ε-approximants, recently introduced by Bukh and Nivasch [5]: For a range space (X, R), a *one-sided ε-approximant* on $P \subset X$ is a subset N of P with the property that for every $r \in R$ we have $\frac{|r \cap P|}{|P|} - \frac{|r \cap N|}{|N|} \leq \varepsilon$. Once again, the restriction that N has to be a subset of P can be dropped to get the notion of weak one-sided ε-approximations. Note that the only difference to the definition of ε-approximations is that one does not take the absolute value of the difference. In particular, if $\frac{|r \cap N|}{|N|} > \frac{|r \cap P|}{|P|}$, i.e., if r contains many points of N despite containing only few points of P, the difference is negative, and thus smaller than ε.

In their paper, Bukh and Nivasch [5] consider the range space of convex sets. They show that any point set in $\mathbb{R}^d$ allows a one-sided ε-approximant for convex ranges of size $g(\varepsilon, d)$, where $g(\varepsilon, d)$ only depends on ε and d, but not on the size of P.

In a similar reasoning, it makes sense to define an approximation by a set N such that for every $r \in R$ we have $\frac{|r \cap N|}{|N|} - \frac{|r \cap P|}{|P|} \leq \varepsilon$. Intuitively, if a range r contains a large fraction of the points of N, then it is guaranteed to contain a large fraction of the set P we want to approximate. But here again, our approximation ratio is $\frac{1}{2}$ at best.

2 Two points

We first consider the case where the underlying data is a point set. Motivated by the definition of generalized Tukey depth, we consider $\alpha_1 = \frac{1}{5}$ and $\alpha_2 = \frac{2}{5}$. Even though this result is a special case of Theorem 1, we still show its proof for two reasons: first, the Algorithm presented in Section 5 relies heavily on the presented proof and, secondly, the proof already illustrates the main ideas for the proof of Theorem 1.

▶ **Theorem 3.** *Let P be a set of n points in general position in the plane. Then there are two points p_1 and p_2 in $\mathbb{R}^2$ such that*

(1) *each closed halfplane containing one of the points p_1 and p_2 contains at least $\frac{n}{5}$ of the points of P and*

(2) *each closed halfplane containing both p_1 and p_2 contains at least $\frac{2n}{5}$ of the points of P.*

Proof. Note that condition (1) is equivalent to the condition that every open halfplane containing more than $\frac{4n}{5}$ of the points of P must contain both p_1 and p_2. Similarly, condition (2) is equivalent to the condition that every open halfplane containing more than $\frac{3n}{5}$ of the points of P must contain one of p_1 and p_2. We will now construct two points p_1 and p_2 satisfying both these conditions.

Let C be the intersection of all open halfplanes containing more than $\frac{4n}{5}$ of the points of P. Clearly C is convex. Also, note that C is closed. The centerpoint region is a strict subset of C and thus C has a non-empty interior. In order to satisfy condition (1), both p_1 and p_2 have to be placed in C.

Let now H be the set of all open halfplanes containing more than $\frac{3n}{5}$ of the points of P. For any h_i in H let c_i be the intersection of h_i and C. In order to also satisfy condition (2), we need to find two points p_1 and p_2 such that every c_i contains at least one of them. To this end, we partition H into two subsets L and R. The set L contains all halfplanes that lie on the left side of their respective boundary lines. Analogously, R contains all halfplanes that lie on the right side of their respective boundary lines. For a halfplane h_i that has a horizontal boundary line, we put h_i in L if and only if it lies above its boundary line.

Note that any three halfplanes in L have a non-empty intersection: Consider the inclusion-minimal halfplane $h \in L$ with horizontal boundary line and its intersection r with the boundary of the convex hull of P. As h is open, r is not in h. However, we claim that any point r' in h on the convex hull boundary of P in an ε-neighborhood of r is in any halfplane of L. Indeed, if there was a halfplane in L not containing r', it would contain a strict subset of the intersection of the convex hull of P with h; however, this would contradict the minimality of h. The analogous holds for R.

We will now show that for any two halfplanes h_1 and h_2 in L, their corresponding regions c_1 and c_2 have a non-empty intersection. The same arguments hold for any two halfplanes in R. Assume for the sake of contradiction that c_1 and c_2 do not intersect. As C and $h_1 \cap h_2$ are convex, this means that there is an open halfplane g containing more than $\frac{4n}{5}$ of the points of P such that the intersection of the boundary lines of h_1 and h_2 lies in $\overline{g}$, the complement of g (see Figure 1). In particular, $g \cap h_1$ is a strict subset of $\overline{h_2}$. As $\overline{g}$ contains strictly fewer than $\frac{n}{5}$ of the points of P and $\overline{h_1}$ contains strictly fewer than $\frac{2n}{5}$ of the points of P, $g \cap h_1$ must contain strictly more than $\frac{2n}{5}$ of the points of P. However, being a subset of $\overline{h_2}$, which also contains strictly fewer than $\frac{2n}{5}$ of the points of P, this is a contradiction. Thus, by contradiction, c_1 and c_2 intersect.

As neither three halfplanes in L nor two halfplanes in L and C have an empty intersection, Helly's Theorem entails that there exists a point in both C and all halfplanes in L, i.e., all c_is associated to L have a non-empty intersection D_L. Again, the same holds for R, with a non-empty intersection D_R. Placing p_1 in D_L and p_2 in D_R, we have thus constructed two points such that the conditions (1) and (2) hold. ◀

This result is tight in the following sense: There is a point set for which it is not possible to improve both conditions at the same time, that is, it is not possible to find two points such that any halfplane containing one of them contains strictly more than $\frac{n}{5}$ of the points and any halfplane containing both of them contains strictly more than $\frac{2n}{5}$ of the points. For

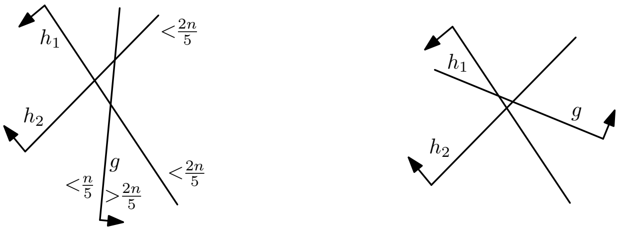

Figure 1 Two c_is associated to L must intersect (left). The intersection is non-empty in other variants (right).

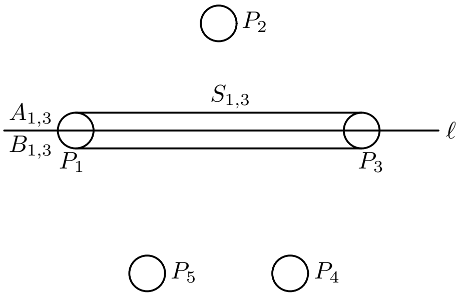

Figure 2 A construction for which the bounds of Theorem 3 cannot be improved.

this consider a set of $n = 5k$ point arranged in the following way. Partition the points into 5 sets $P_1, \ldots, P_5$ of k points each. Place $P_1, \ldots, P_5$ in such a way that the convex hull of each P_i is disjoint from the convex hull of the union of the other four sets (see Figure 2).

Denote by $S_{i,j}$ the convex hull $\mathsf{CH}(P_i \cup P_j)$ of $P_i \cup P_j$. Let ℓ be a line through $\mathsf{CH}(P_i)$ and $\mathsf{CH}(P_j)$. Note that any other set P_m is not separated by ℓ (i.e., lies entirely on one side). Let $A_{i,j}$ be the side of ℓ containing fewer of the other sets and let $B_{i,j}$ be the other side. For any point q in $\mathsf{CH}(P_1 \cup \ldots \cup P_5)$ we say that q is *above* $S_{i,j}$ if it is not in $S_{i,j}$ but it is in $A_{i,j}$. Similarly, for any point q in $\mathsf{CH}(P_1 \cup \ldots \cup P_5)$ we say that q is *below* $S_{i,j}$ if it is not in $S_{i,j}$ but it is in $B_{i,j}$. Suppose, for the sake of contradiction, that there exist two points p_1 and p_2 such that any halfplane containing one of them contains strictly more than k of the points of $P_1 \cup \ldots \cup P_5$ and any halfplane containing both of them contains strictly more than $2k$ of the points of $P_1 \cup \ldots \cup P_5$. Consider two sets P_i and P_j such that $A_{i,j}$ contains exactly one other set. First we note that neither p_1 nor p_2 can lie above $S_{i,j}$ as otherwise we can find a halfplane containing that point and only one of the sets, i.e., only k points. Similarly, we cannot place both p_1 and p_2 below $S_{i.j}$, as otherwise we can find a halfplane containing both points and only two of the sets, i.e., only $2k$ points. Also, we must clearly place both p_1 and p_2 in $\mathsf{CH}(P_1 \cup \ldots \cup P_5)$. Thus, for any two sets P_i and P_j such that $A_{i,j}$ contains exactly one other set, $S_{i,j}$ must contain at least one of p_1 and p_2. However, there are five such $S_{i,j}$ and $P_1, \ldots, P_5$ can be placed in such a way that no three of them have a common intersection. So no matter how we place p_1 and p_2, one of the $S_{i,j}$ will be empty.

3 An arbitrary number of points

We now strengthen Theorem 3 in four ways: we allow for arbitrarily many query points, we extend it to higher dimensions, we consider mass distributions instead of point sets, and we give a range of possible bounds:

▶ **Theorem 1.** *Let μ be a mass distribution in $\mathbb{R}^d$ with $\mu(\mathbb{R}^d) = 1$. Let $\alpha_1, \ldots, \alpha_k$ be non-negative real numbers such that $\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_k$ and for every i, j with $i + j \leq k + 1$ we have that $(d-1)\alpha_k + \alpha_i + \alpha_j \leq 1$. Then there are k points $p_1, \ldots, p_k$ in $\mathbb{R}^d$ such that for each closed halfspace h containing j of the points $p_1, \ldots, p_k$ we have $\mu(h) \geq \alpha_j$.*

We use the following observation, which follows from the fact that for an empty intersection of $d + 1$ halfspaces, any point with non-zero mass is in at most d such halfspaces.

▶ **Observation 4.** *Let μ be a mass distribution in $\mathbb{R}^d$ with $\mu(\mathbb{R}^d) = 1$. Let $h_1, \ldots, h_{d+1}$ be $d + 1$ open halfspaces with $h_1 \cap \ldots \cap h_{d+1} = \emptyset$. Then $\mu(h_1) + \ldots + \mu(h_{d+1}) \leq d$.*

Proof of Theorem 1. The result is straightforward for $d = 1$, so assume $d \geq 2$. Again the condition that for each closed halfspace h' containing j of the points $p_1, \ldots, p_k$ we have $\mu(h') \geq \alpha_j$ is equivalent to the condition that every open halfspace h with $\mu(h) > 1 - \alpha_j$ must contain at least $k - j$ of the points $p_1, \ldots, p_k$. Let $\alpha_0 = 0$. For $1 \leq j \leq k$, we call an open halfspace h a *j-halfspace* if $1 - \alpha_{k-j+1} < \mu(h) \leq 1 - \alpha_{k-j}$. Consider the x_1-x_2-plane, denoted by X, and for each vector $v = (v_1, v_2, \ldots, v_d)$ in $\mathbb{R}^d$ let $\pi(v) = (v_1, v_2, 0, \ldots, 0)$ be the projection of v to X. Let $v_1, \ldots, v_k$ be k unit vectors in X with the property that the angle between any v_i and v_{i+1} is $\frac{2\pi}{k}$. Note that this implies that also the angle between v_k and v_1 is $\frac{2\pi}{k}$. For each v_i we construct a *principal set* V_i of halfspaces as follows: For each j, consider all j-halfspaces. For any such halfspace h, let $n(h)$ be the normal vector to its bounding hyperplane that points into h. Let h be in V_i if the angle between $\pi(n(h))$ and v_i is at most $\frac{j\pi}{k}$. If $\pi(n(h)) = 0$, place h arbitrarily in j of the V_i's. Note that with this construction each j-halfspace is contained in exactly j principal sets. Thus, if, for each principal set, we can pick a point in all its halfplanes, then each j-halfplane contains j points.

It remains to show that the halfspaces in each principal set have a common intersection. Let $h_1, \ldots, h_{d+1}$ be $d + 1$ halfspaces in V_i and assume for the sake of contradiction that they have no common intersection. Then the positive hull (conical hull) of their projected normal vectors must be X, and in particular there are three of them, w.l.o.g. h_1, h_2 and h_3, whose projected normal vectors already have X as their positive hull. Further, among those three halfspaces, there are two of them, w.l.o.g. h_1 and h_2, such that the angles between their projected normal vectors and v_i sum up to more than π. If h_1 is a j_1-halfspace, then by construction of V_i we have that the angle between $\pi(n(h_1))$ and v_i is at most $\frac{j_1\pi}{k}$. Analogously, if h_2 is a j_2-halfspace, the angle between $\pi(n(h_2))$ and v_i is at most $\frac{j_2\pi}{k}$. By the choice of h_1 and h_2 we thus have $\frac{(j_1+j_2)\pi}{k} > \pi$, which is equivalent to $j_1 + j_2 > k$, and to $j_1 + j_2 \geq k + 1$, as j_1 and j_2 are integers. By definition of a j-halfspace we have

$$\mu(h_1) + \mu(h_2) > 1 - \alpha_{k+1-j_1} + 1 - \alpha_{k+1-j_2} \quad .$$

Furthermore we have $\mu(h_i) > 1 - \alpha_k$ for every $i \in \{1, \ldots, d + 1\}$, and thus

$$\mu(h_1) + \mu(h_2) + \mu(h_3) + \ldots + \mu(h_{d+1}) > 1 - \alpha_{k+1-j_1} + 1 - \alpha_{k+1-j_2} + (d-1)(1 - \alpha_k) \quad ,$$

which is equivalent to

$$(d-1)\alpha_k + \alpha_{k+1-j_1} + \alpha_{k+1-j_2} > d + 1 - (\mu(h_1) + \ldots + \mu(h_{d+1})) \quad .$$

As $k+1-j_1+k+1-j_2 = 2k+2-(j_1+j_2) \leq k+1$, we have that $(d-1)\alpha_k+\alpha_{k+1-j_1}+\alpha_{k+1-j_2} \leq 1$ and thus $\mu(h_1) + \ldots + \mu(h_{d+1}) > d$, which is a contradiction to Observation 4. ◀

Setting $\alpha_j = \frac{j}{kd+1}$, we get a bound for the generalized Tukey depth:

▶ **Corollary 5.** *Let μ be a mass distribution in $\mathbb{R}^d$ with $\mu(\mathbb{R}^d) = 1$. Then there exist k points $p_1, \ldots, p_k$ in $\mathbb{R}^d$ with generalized Tukey depth $\mathsf{gtd}_\mu(\{p_1, \ldots, p_k\}) = \frac{1}{kd+1}$.*

4 Triangles

As mentioned before, the $\frac{1}{3}$-quantile and the $\frac{2}{3}$-quantile can also be interpreted as a one-dimensional simplex with the property that every halfline that contains a part of the simplex contains at least $\frac{1}{3}$ of the underlying data set and every halfline that contains the whole simplex contains at least $\frac{2}{3}$ of the underlying data set. For this interpretation, we give a generalization to two dimensions. For ease of presentation, we only give a proof for point sets instead of mass distributions and for fixed values of α and β.

▶ **Theorem 6.** *Let P be a set of n points in general position in the plane. Then there are three points p_1, p_2 and p_3 in $\mathbb{R}^2$ such that*

(1) *each closed halfplane containing one of the points p_1, p_2 and p_3 contains at least $\frac{n}{6}$ of the points of P and*

(2) *each closed halfplane containing all of p_1, p_2 and p_3 contains at least $\frac{n}{2}$ points of P.*

Note that this can also be interpreted as an instance of Theorem 1 with $\alpha_1 = \alpha_2 = \frac{1}{6}$ and $\alpha_3 = \frac{1}{2}$. However, as $\alpha_3 + \alpha_3 + \alpha_1 > 1$, the precondition of Theorem 1 does not apply.

As the proof of this result uses similar ideas as the above proofs, we only sketch the main ideas and refer the interested reader to the full version.

Sketch of proof. Let C be the intersection of all open halfplanes containing more than $\frac{5n}{6}$ of the points of P. Just as in the proof of Theorem 3, condition (1) is equivalent to p_1, p_2 and p_3 lying in C. Similarly, condition (2) is equivalent to the following statement: for every halfplane h containing more than $\frac{n}{2}$ of the points of P, h contains at least one of p_1, p_2 and p_3. For each such h, let c_h be the intersection of h and C and let H be the set of all c_h's that are minimal with respect to inclusion. It can be shown that among any three elements of H, two of them intersect. Using this property, we can then place 3 points on the boundary of C such that each element of H contains at least one of them: Place p_1 at a topmost point of the boundary of C. Let h_1 be the first element of H in counterclockwise direction whose defining halfplane does not contain p_1. Place p_2 at the intersection of the defining line of h_1 with the boundary of C that is furthest in counterclockwise direction from p_1. Since h_1 is minimal, any element of H intersecting h_1 contains either p_1 or p_2. Further, all elements of H that do not intersect h_1 have a common intersection, in which we place p_3. ◀

The general statement can be proved analogously:

▶ **Theorem 2.** *Let μ be a mass distribution in $\mathbb{R}^2$ with $\mu(\mathbb{R}^2) = 1$. Let α and β be real numbers such that $0 < \alpha \leq \beta$ and $\alpha + \beta = \frac{2}{3}$. Then there is a triangle Δ in $\mathbb{R}^2$ such that*

(1) *for each closed halfplane h containing one of the vertices of Δ we have $\mu(h) \geq \alpha$ and*

(2) *for each closed halfplane h fully containing Δ we have $\mu(h) \geq \beta$.*

5 Construction in the plane

In this section, we describe algorithms for constructing the points described in Theorems 3 and 6. We first observe that the convex regions defined by the intersections of the half-planes in sets like L and R in the proof of Theorem 3 correspond to levels in the dual line arrangement. We use the duality $p^* = (y = kx + d) \iff p = (k, d)$ that maps a point p to a line p^*. The *k-level* of a line arrangement is the set of points with exactly $k - 1$ lines below it and not more than $n - k$ lines above it. (It thus consists of segments of the line arrangement.) Suppose we are given α_1 and α_2, s.t. $0 < \alpha_1 \leq \alpha_2$ and $\alpha_1 + 2\alpha_2 = 1$. Let U be the set of open halfplanes that are above their boundary lines and contain more than

$(1-\alpha_2)n$ points of P, and let D_U be their intersection. A point p is in D_U if there is no line through it having at least $\lfloor(1-\alpha_2)n+1\rfloor$ points of P above it. If the dual line p^* of p contains a point ℓ^* below the $\lceil\alpha_2 n\rceil$-level of the dual line arrangement of P, then p has a supporting line ℓ with more than $(1-\alpha_2)n$ points of P above it. Since a line has a point below that level if and only if it intersects the interior of its convex hull, the interior of the convex hull of the $\lceil\alpha_2 n\rceil$-level thus excludes exactly those lines whose primal points are not in D_U. The supporting lines of the segments of the convex hull of the $\lceil\alpha_2 n\rceil$-level give the primal points that bound D_U. Matoušek [10] describes an algorithm for constructing the k-level of a line arrangement in $O(n\log^4 n)$ time. The *k-hull* of a set P of n points in the plane is the set of points p in $\mathbb{R}^2$ such that any closed halfplane defined by a line through p contains at least k points of P. The set C in the proof of Theorem 3 is the intersection of all open halfplanes containing more than $\frac{4n}{5}$ points. C is thus the $\lceil\frac{n}{5}\rceil$-hull of P. The k-hull of P is obtained by computing the convex hulls of the k-level and the $(n-k)$-level of the dual line arrangement of P, which give the upper and lower envelope of the k-hull [10]. To construct the points from Theorems 3 and 6 (without explicitly constructing the levels), we use Matoušek's algorithmic tools from [10]. (Alternatively, a general optimization technique by Langerman and Steiger [9] can be used, as detailed in the full version.)

▶ **Lemma 7** (Matoušek [10, Lemma 3.2]). *In an arrangement of n lines, let γ be the boundary of the convex hull of the lines on or below the k-level. Given the arrangement, k, and a point p, one can find the tangent to γ passing through p and touching γ to the right of p (if it exists) in time $O(n\log^2 n)$.*

▶ **Lemma 8.** *Given an arrangement of n lines and two numbers $k < l \leq n$, as well as a halfplane h, a line separating the k-level from the intersection of h with the l-level can be found in $O(n\log^3 n)$ time, if it exists. The separating line is tangent to both level parts and, from left to right, first intersects the k-level and then the relevant part of the l-level.*

Proof. Let γ be the boundary of the convex hull of all points below the k-level, and let ν be the intersection of h with the l-level. Note that ν might not be connected. Suppose we want our line to be the counterclockwise bitangent of γ and ν (i.e., from left to right, it first intersects γ, which has no point above it, and then ν). Our algorithm works by obtaining tangents to ν through points on γ. Matoušek's $O(n\log^2 n)$ algorithm for determining the tangent to a level through a given point that is to the right of that point [10, Lemma 3.2] (our Lemma 7) also directly works for parts of a level such as ν: It requires a sub-algorithm that decides in $O(n\log n)$ time whether a given line ℓ intersects the level (or, in our case, the partial level ν). This can be done by sorting the intersection of the lines of the arrangements along ℓ (see also [10, Lemma 3.1]) as well as along the line bounding h; ℓ either intersects the relevant part of ν, or we can compare the intersection of h with ℓ to the intersections of h with ν to determine whether there is a point of ν below ℓ.

Suppose first we are given γ. (It requires $O(n\log^4 n)$ time though to obtain it, so we eventually get rid of this assumption.) The convex hull of a level is known to have at most n vertices [10, Lemma 2.1]. For a point p on γ, we can find in $O(n\log^2 n)$ time the point q on ν such that the line pq has no point on ν below it. We can thus find, by binary search on the $O(n)$ vertices of γ, a vertex p with q on ν such that pq separates γ and ν. This gives an $O(n\log^4 n)$ time algorithm for obtaining the bitangent. To improve on that bound, we need to get rid of the explicit construction of γ to find the tangents to ν.

To this end, we consider Matoušek's algorithm for constructing the convex hull boundary γ of a level and compute only the relevant part (see [10, Section 4]). In particular, the algorithm works by finding, for a constant c and two vertical lines, $(c-1)$ further vertical lines between

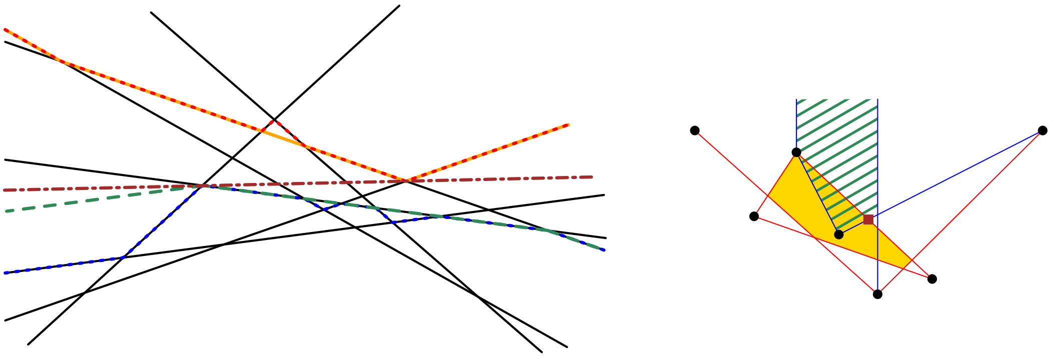

Figure 3 A counterclockwise bitangent (brown, dash-dotted) between the $\lceil \frac{2n}{5} \rceil$-level (blue) and the $\lfloor \frac{4n}{5} \rfloor$-level (red) of an arrangement of seven lines (left). The primal point configuration is shown to the right; there, the orange region corresponds to the $\lceil \frac{n}{5} \rceil$-hull C, and the hatched green region corresponds to D_U. Observe that there can be vertices of D_U outside of C.

the given ones such that there are at most n^2/c crossings of the arrangement between two of these verticals. This can be done in $O(n)$ time (as described in [11]). The tangents on γ at the intersection points with the vertical lines can be computed in $O(n \log^3 n)$ time [10, Lemma 3.3]. It is shown in [10] that, when choosing $c = 64$, there are at most $n/2$ lines of the arrangement relevant for the construction of γ between two such vertical lines, and these lines can be found in $O(n)$ time. The original algorithm proceeds recursively within each interval defined by two neighboring vertical lines after removing the non-relevant lines. In our adaption, however, we find the interval that contains the point p on γ such that a tangent to γ through the vertex p with q on ν such that pq separates γ and ν. (We do this by considering the tangent to γ at each of the constant number of intersection of a vertical line with γ.) When we have found this interval, we can prune $n/2$ of the lines and recurse inside this interval. Note, however, that we cannot prune the set of lines when looking for a tangent to ν. Thus, in each recursive call, we need $O(n \log^2 n)$ time for computing the tangent. As the recursion depth is $O(\log n)$, this amounts to $O(n \log^3 n)$ in total. Also, for n_i lines during the ith recursion, we need $O(n_i \log^3 n_i) \subseteq O(n_i \log^3 n)$ time for determining the intervals. As n_i decreases geometrically, this also amounts to $O(n \log^3 n)$. This is the total running time for finding the bitangent, as claimed. ◀

We call such a line the *counterclockwise bitangent* of the two subsets of the plane (i.e., the intersection with the region not above it has smaller x-coordinate than the intersection with the region not below it). Note that by mirroring the plane horizontally or vertically, the lemma also provides other types of bitangents. Figure 3 shows an example.

▶ **Theorem 9.** *Given a set P of n points in the plane, two points satisfying the conditions of Theorem 3 can be constructed in time $O(n \log^3 n)$.*

Proof. To find a point p_1 in the intersection of C and D_U, observe first that we can restrict our attention in the dual to the convex hull of the points above the $\lfloor (1-\alpha_1)n \rfloor$-level of the dual line arrangement. This is because any primal line with more than $(1-\alpha_1)n$ points above it (which corresponds to a dual point below the $\lceil \alpha_1 n \rceil$-level) also defines a halfplane in U. A point in the intersection of D_U and C thus corresponds to a line on or above the $\lceil \alpha_2 n \rceil$-level and on or below the $\lfloor (1-\alpha_1)n \rfloor$-level. We find a bitangent to these two levels in $O(n \log^3 n)$ time using Lemma 8 (with $h = \mathbb{R}^2$). The primal point of this line is p_1; see the point indicated by the brown box in Figure 3 (right). We obtain p_2 analogously. ◀

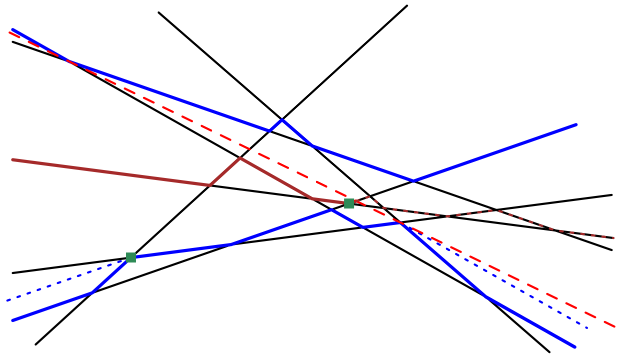

Figure 4 An arrangement of seven lines with the $\lceil \frac{n}{6} \rceil$-level and $\lfloor \frac{5n}{6} \rfloor$-level (blue) and the clockwise bitangent p_1^* (red dashed) between them. The green boxes indicate the two points defining the counterclockwise bitangent between the $\lceil \frac{n}{6} \rceil$-level and μ_1 (brown).

▶ **Theorem 10.** *Three points as described in Theorem 6 can be computed in time $O(n \log^3 n)$.*

Proof. Consider the dual line arrangement of the point set. The points p_1, p_2, p_3 dualize to three lines p_1^*, p_2^*, p_3^* that are between the $\lceil \frac{n}{6} \rceil$-level and the $\lfloor \frac{5n}{6} \rfloor$-level of the arrangement s.t. every point on the middle level has at least one of these lines above it and one of these lines below it. (We assume for simplicity that n is odd and the *middle level* is the $\lfloor \frac{n}{2} \rfloor$-level of the arrangement; if n is even, one has to consider the points between the $\frac{n}{2}$-level and the $(\frac{n}{2}+1)$-level.) Theorem 6 asserts that such lines exist, and its proof tells us that we can choose one of these lines to be an arbitrary tangent of one of the levels not intersecting the interior of the other one. We denote by γ_b and γ_t the convex hull boundaries of the points on or below the $\lceil \frac{n}{6} \rceil$-level and of the points on or above the $\lfloor \frac{5n}{6} \rfloor$-level, respectively.

We let p_1^* be the clockwise bitangent of γ_b and γ_t, which we can obtain in $O(n \log^3 n)$ time using Lemma 8. For simplicity of explanation, we also compute the counterclockwise bitangent ℓ. (This step may be omitted in an actual implementation, but assuming it to be given facilitates the explanation and does not change the asymptotic running time.)

The line p_1^* intersects the middle level of the arrangement. Let μ_1 be the parts of the middle level below p_1^*, and μ_2 be the part above it. Note that each of these parts may be disconnected. Using Lemma 8, we search for the counterclockwise bitangent between γ_b (or, equivalently, the $\lceil \frac{n}{6} \rceil$-level) and μ_1 (which is the intersection of the middle level with a halfspace defined by p_1^*) in $O(n \log^3 n)$ time. If it exists, and its intersection point with γ_b is between the intersections of γ_b with p_1^* and ℓ, we choose this line to be p_2^*. Otherwise, we continue our search on γ_t) in the same way (i.e., we look for the counterclockwise bitangent between γ_t and μ_1). The line p_3^* can be found in an analogous manner. ◀

6 Conclusion

We proposed a generalization of quantiles in higher dimensions based on a generalization of Tukey depth to multiple points. Our bounds and algorithms seem merely being a first step in this direction and we can identify several interesting open problems. Except for special cases of Theorem 1, we do not believe that our bounds are tight and particularly expect significantly better bounds in higher dimensions. Naturally, there are many other range spaces for which this problem could be considered, e.g., convex sets, like in [5].

From an algorithmic point of view, the bottleneck for the running time of our approach is Lemma 8. The current methods result in $O(n \log^3 n)$ time. While solutions to such kinds of problems can usually only be verified in $\Theta(n \log n)$ time (see, e.g., [2, 16]), a linear-time algorithm, like for centerpoints [8], is conceivable. For arbitrarily many points, it seems tedious but doable to apply similar approaches as in the proof of Theorem 9. Is there a good bound on the running time independent of the size of $|Q|$?

References

1 Greg Aloupis. Geometric measures of data depth. In Regina Y. Liu, Robert Serfling, and Diane L. Souvaine, editors, *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications*, pages 147–158. DIMACS/AMS, 2003.

2 Greg Aloupis, Carmen Cortés, Francisco Gómez, Michael Soss, and Godfried Toussaint. Lower bounds for computing statistical depth. *Comput. Statist. Data Anal.*, 40(2):223–229, 2002. doi:10.1016/S0167-9473(02)00032-4.

3 Boris Aronov, Franz Aurenhammer, Ferran Hurtado, Stefan Langerman, David Rappaport, Carlos Seara, and Shakhar Smorodinsky. Small weak epsilon-nets. *Comput. Geom.*, 42(5):455–462, 2009. doi:10.1016/j.comgeo.2008.02.005.

4 Maryam Babazadeh and Hamid Zarrabi-Zadeh. Small Weak Epsilon-Nets in Three Dimensions. In *Proc. 18th Canadian Conference on Computational Geometry (CCCG)*, 2006. URL: http://www.cs.queensu.ca/cccg/papers/cccg13.pdf.

5 Boris Bukh and Gabriel Nivasch. One-Sided Epsilon-Approximants. In Martin Loebl, Jaroslav Nešetřil, and Robin Thomas, editors, *A Journey Through Discrete Mathematics: A Tribute to Jiří Matoušek*, pages 343–356. Springer, 2017. doi:10.1007/978-3-319-44479-6_12.

6 Probal Chaudhuri. On a geometric notion of quantiles for multivariate data. *J. American Statist. Assoc.*, 91(434):862–872, 1996.

7 David Haussler and Emo Welzl. ε-nets and simplex range queries. *Discrete Comput. Geom.*, 2:127–151, 1987. doi:10.1007/BF02187876.

8 Shreesh Jadhav and Asish Mukhopadhyay. Computing a Centerpoint of a Finite Planar Set of Points in Linear Time. *Discrete Comput. Geom.*, 12:291–312, 1994. doi:10.1007/BF02574382.

9 Stefan Langerman and William L. Steiger. Optimization in Arrangements. In *20th Symp. on Theoretical Aspects of Computer Science (STACS)*, volume 2607 of *LNCS*, pages 50–61, 2003. doi:10.1007/3-540-36494-3_6.

10 Jiří Matoušek. Computing the Center of Planar Point Sets. In Jacob E. Goodman, Richard Pollack, and William Steiger, editors, *Discrete and Computational Geometry: Papers from the DIMACS Special Year*, volume 6 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 221–230. DIMACS/AMS, 1990.

11 Jiří Matoušek. Construction of epsilon-Nets. *Discrete Comput. Geom.*, 5:427–448, 1990. doi:10.1007/BF02187804.

12 Jiří Matoušek. Tight upper bounds for the discrepancy of half-spaces. *Discrete Comput. Geom.*, 13(3):593–601, 1995. doi:10.1007/BF02574066.

13 Jiří Matoušek, Emo Welzl, and Lorenz Wernisch. Discrepancy and approximations for bounded VC-dimension. *Combinatorica*, 13(4):455–466, 1993. doi:10.1007/BF01303517.

14 Nabil Mustafa and Kasturi Varadarajan. Epsilon-approximations and epsilon-nets. In *Handbook of Discrete and Computational Geometry.* HAL, 2017. URL: https://hal.archives-ouvertes.fr/hal-01468664.

15 Nabil H. Mustafa and Saurabh Ray. An optimal extension of the centerpoint theorem. *Comput. Geom.*, 42(6):505–510, 2009. doi:10.1016/j.comgeo.2007.10.004.

16 Sambuddha Roy and William Steiger. Some Combinatorial and Algorithmic Applications of the Borsuk-Ulam Theorem. *Graphs Combin.*, 23:331–341, 2007. doi:10.1007/s00373-007-0716-1.

17 Mudassir Shabbir. *Some results in computational and combinatorial geometry.* PhD thesis, Rutgers The State University of New Jersey, 2014.

18 John W. Tukey. Mathematics and the picturing of data. In *Proc. International Congress of Mathematicians*, pages 523–531, 1975.