Computational Geometry: Theory and Applications 103 (2022) 101856

Contents lists available at ScienceDirect

Computational Geometry: Theory and Applications

www.elsevier.com/locate/comgeo

Expected size of random Tukey layers and convex layers

Zhengyang Guo [a,*], Yi Li [a], Shaoyu Pei [b]

[a] School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, Singapore
[b] College of Science and Mathematics, California State University, Fresno, CA, United States

ARTICLE INFO

Article history:
Received 20 March 2020
Received in revised form 19 July 2021
Accepted 12 December 2021
Available online 21 December 2021

Keywords:
Convex layer
Tukey depth
Tukey layer
Computational geometry
Geometric probability

ABSTRACT

We study the Tukey layers and convex layers of a planar point set, which consists of n points independently and uniformly sampled from a convex polygon with k vertices. We show that the expected number of vertices on the first t Tukey layers is $O(kt\log(n/k))$ and the expected number of vertices on the first t convex layers is $O\left(kt^3\log(n/(kt^2))\right)$. We also show a lower bound of $\Omega(t\log n)$ for both quantities in the special cases where $k=3,4$. The implications of those results in the average-case analysis of two computational geometry algorithms are then discussed.

1. Introduction

The motivation of this work is to understand the combinatorial and geometric properties of random convex layers and Tukey layers of planar point sets X. The convex layers of X are a sequence of nested convex polygons whose vertices form a partition of X. The Tukey layers are the cells of a partition of X, in which each cell consists of all points in X of the same Tukey depth [1]. We refer the readers to Definitions 1 and 4 for precise definition. Each Tukey layer, as we shall prove in Lemma 1, is exactly the vertices of a convex polygon.

There has been a long research history on the expected size of the convex hull of a random point set [2–5], the relation between the expected size and the expected area of the convex hull [6,7], and the expected convex depth [8]. However, few of them study convex layers. In fact, the vertices on the first t convex layers, denoted by $V_{[t]}(X)$, are closely related to the partial enclosing problem introduced by Atanassov et al. in [9]. The objective of this problem is to find the convex hull with the minimum area that encloses $(n-t)$ of the n points in X. The t excluded points are regarded as outliers, as in many works that study the partial covering, for example [10], [11] and [12].

In [9], Atanassov et al. give an algorithm with the worst-case time complexity of $O\left(n\log n+\binom{4t}{2t}(3t)^t n\right)$, where the n in the second term $\binom{4t}{2t}(3t)^t n$ refers to the size $\left|V_{[t]}(X)\right|$ in the worst case. However, the actual runtime seldom meets such worst cases. To give an overall measure on the efficiency of the algorithm, it makes more sense to study the average time complexity. Assuming that X is uniformly sampled from a convex k-gon as in [2,13,7,6,14,5,15], we shall prove in Section 4 that $\mathbb{E}\left|V_{[t]}(X)\right|=O(kt^3\log(n/(kt^2)))$, which is $o(n)$ when $t=o((n/(k\log(nk))^{1/3})$. As a consequence, the expected

* Corresponding author.
E-mail addresses: GUOZ0015@e.ntu.edu.sg (Z. Guo), yili@ntu.edu.sg (Y. Li), Shaoyupei@mail.fresnostate.edu (S. Pei).

https://doi.org/10.1016/j.comgeo.2021.101856
0925-7721/

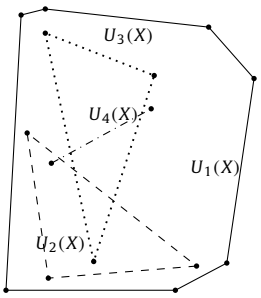

Fig. 1. The boundary of the first three Tukey layers $U_1(X)$, $U_2(X)$ and $U_3(X)$ is plotted in solid, dashed, and dotted lines, respectively. The fourth Tukey layer $U_4(X)$ degenerates to a line segment, plotted in dashed dots. The vertices in each Tukey layer are in the convex positions.

complexity of Atanassov et al.'s algorithm in [9] is $O(n\log n + \binom{4t}{2t}(3t)^t kt^3\log(n/t^2))$. This explains the gap between the worst case complexity and the actual runtime.

In addition, we study the expected number of vertices on the first t Tukey layers $U_{[t]}(X)$ as defined in Definition 4. This is also related to a partial shape fitting problem [16] in which the parallelogram rather than the convex polygon as in [9] is concerned. The time complexity of the algorithm in [16] is $O(n^2t^4 + n^2\log n)$, where the n in the first term n^2t^4 refers to $\left|U_{[t]}(X)\right|$ in the worst case. As we shall prove $\mathbb{E}\left|U_{[t]}(X)\right| = O(kt\log(n/k))$ in Section 3, the expected time complexity is then $O\left(kt^5 n\log(n/k) + n^2\log n\right)$, smaller than the worst-case complexity when $\Omega\left((n/k)^{1/5}\right) \le t \le O(n/(k\log n))$.

It is beneficial to study the convex hulls and Tukey layers together. Their close relation is shown in Lemma 2 that $U_{[t]}(X) \subseteq V_{[t]}(X)$. An upper bound on $\left|V_{[t]}(X)\right|$ is then automatically an upper bound on $\left|U_{[t]}(X)\right|$ and a lower bound on $\left|U_{[t]}(X)\right|$ is automatically a lower bound on $\left|V_{[t]}(X)\right|$.

1.1. Notation and definitions

We introduce the notation and definitions before reviewing the existing works. Let X be a planar point set and $n = |X|$ be its size. When X is a random point set, we use $\mathcal{P}$ to denote the convex polygon from which X is sampled, and k to denote the number of vertices of $\mathcal{P}$. Throughout this work, the convex polygon $\mathcal{P}$ is always closed and, without loss of generality, we assume the area of $\mathcal{P}$ is 1. We now present the definition of the convex layer structure as in [17].

Definition 1 *(Convex layer).* Given a planar point set X, the first convex layer $H_1(X)$ is defined to be the convex hull $H(X)$ of the whole point set. The t-th convex layer $H_t(X)$ is inductively defined to be the convex hull of the remaining points, after the points on the first $(t-1)$ convex layers have been removed from X.

Definition 2 *(Convex depth).* The convex depth of $p \in X$ is said to be t if p is a vertex of $H_t(X)$.

Next we define the Tukey layers, for which we need to introduce a classical notion known as the Tukey depth [1]. Instead of using the original definition, we use the following equivalent form for finite point sets.

Definition 3 *(Tukey depth).* Given a set X of planar points, the Tukey depth of a point $p \in X$ is defined to be $N(p)+1$, where $N(p)$ is the minimum number of points in X that are contained in any open half-plane with p on its boundary.

Remark 1. For brevity, we use "one side of a line ℓ" to refer to one of the two open half-planes induced by ℓ. Hence, if a point p is on one side of a line ℓ, the point p is in an open half-plane induced by ℓ. Besides, when we say a point is above (below) a line, we do not include the line either.

Remark 2. Intuitively, if a point p has Tukey depth t, then for all lines ℓ through p, there cannot be fewer than $(t-1)$ points on either side of ℓ. At the same time, there exists a line ℓ_0 through p such that there are exactly $(t-1)$ points on one side of ℓ_0.

Definition 4 *(Tukey layer).* For $t \ge 1$, the subset $U_t(X)$ of X is defined to be the set of points of Tukey depth t. The t-th Tukey layer, denoted by $S_t(X)$, is defined to be convex hull of $U_t(X)$. The size of $S_t(X)$ is defined to be $|U_t(X)|$.

An illustration of Tukey layers is shown in Fig. 1. As we shall prove in Lemma 1, the points in $U_t(X)$ are in the convex position and are thus exactly the vertices of $S_t(X)$, hence our definition of the size of $S_t(X)$ makes sense. The frequently used notations are listed in Table 1. Note that $S_1(X) = H(X)$ by definition. For convenience, we also let $V_{[t]}(X) := \bigcup_{i=1}^{t} V_i(X)$ and $U_{[t]}(X) := \bigcup_{i=1}^{t} U_i(X)$.

Table 1
Notations used in this work.

Symbol	Definition	Symbol	Definition
$H(X)$	the convex hull of X	$H_t(X)$	the t-th convex layer of X
$V(X)$	the vertices of $H(X)$	$V_t(X)$	the vertices of $H_t(X)$
		$S_t(X)$	the t-th Tukey layer of X
		$U_t(X)$	the vertices of $S_t(X)$
$A(X)$	the area of $H_1(X)$ or $S_1(X)$	$A_t(X)$	the area of $S_t(X)$

1.2. Related work

The main results in Section 3 and 4 are proved using the techniques developed for computing the expected convex hull size. We thus review the works that study the random convex hull, in terms of its area and the number of its vertices. Most of the research interests have been in their expectations, concentration bounds and asymptotic behaviors.

A fundamental result is that the expected size of a random convex hull is $O(k \log n)$, when a large number n of points are independently and uniformly sampled from a convex k-gon. The result was first stated by Rényi and Sulanke in [2] and a geometric proof was later provided by Har-Peled [5, Section 2]. By the relation $\mathbb{E}\,|V(X)| = n[1 - \mathbb{E}\,A(X)]$ proposed in [13] (the area of the k-gon is assumed to be 1 without loss of generality), an upper bound on $\mathbb{E}\,|V(X)|$ will follow from a lower bound on $\mathbb{E}\,A(X)$. Thus in [5], the effort is devoted to deriving a lower bound on the expected area of the convex hull. A critical observation in [5, Section 2] is that, if $p \in X$ is a vertex of the convex hull, then there exists a line ℓ through p such that one side of ℓ contains no points of X. This gives a necessary condition on $p \in H(X)$, and a lower bound on the probability of the event $p \in H(X)$ can then be obtained. Multiplying this lower bound by n immediately yields an lower bound on $\mathbb{E}\,A(X)$.

In addition, there have been a number of studies on the asymptotic behaviors of the convex hull size, such as [2,6,18–20]. Rényi and Sulanke proved that, given X uniformly sampled from a convex k-gon on a plane, the expected size of the convex hull $\mathbb{E}\,|V(X)|$ is asymptotically $\frac{2}{3}k \log n + O(1)$ as $n \to \infty$, where the constant term depends on the polygon [2]. Affentranger and Wieacker generalized the result to higher dimensions and showed that, given that X is uniformly sampled from a simple polytope in $\mathbb{R}^d$ with k vertices, $\mathbb{E}\,|V(X)| = \frac{d}{(d+1)^{d-1}} k \log^{d-1} n + O(\log^{d-2} n)$ [6]. Masse proved that in the planar case, $|V(X)|/(\frac{2}{3}k \log n)$ converges to 1 in probability [19].

There are also studies that assume different underlying distribution for the point set. When the n points are sampled independently from a coordinate-wise independent distribution in $\mathbb{R}^d$, it is proved by He et al. in [21] that the expected size of the t-th convex layer is $O(t^d \log^{d-1}(n/t^d))$. Some studies assume the point set is sampled independently and uniformly from other shapes rather than a convex polygon. In the case of a disc, the expected size of the convex hull is $\Theta(n^{1/3})$, due to Raynaud [22].

1.3. Our contribution

In this work, we introduce a new definition called Tukey layer and provide some fundamental properties of it. Then we study the expected size of the Tukey layers and convex layers when the point set X is uniformly sampled from a k-gon. We show that the expected number of vertices of the first t Tukey layers $\mathbb{E}\left|U_{[t]}(X)\right| = O(kt \log(n/k))$ and that of the first t convex layers $\mathbb{E}\left|V_{[t]}(X)\right| = O(kt^3 \log(n/kt^2))$. The first work to study the expected size of convex layers is [21] where He et al. proved that $\mathbb{E}\,|V_t(X)| = O(t^2 \log(n/t^2))$ when X follows a continuous component independent distribution. Their result can be extended to the cases when X is sampled from a square or more generally a parallelogram, and their bound $O(t^2 \log(n/t^2))$ is better than ours $O(t^3 \log(n/t^2))$ in such cases. On the other hand, the techniques developed in [21] are towards the continuous component independent distribution, and we find it hard to extend them to other polygonal shapes except square or parallelogram. We also prove a matching lower bound $\mathbb{E}\left|U_{[t]}(X)\right| = \Omega(t \log n)$ when X is sampled from a triangle or a parallelogram, which, since $U_{[t]}(X) \subseteq V_{[t]}(X)$, is also a lower bound for $\mathbb{E}\left|V_{[t]}(X)\right|$ in the two special cases. Finally, we show that the two upper bounds are helpful in understanding the average case complexity of two partial shape fitting algorithms, both of which aim to enclose $(n-t)$ of the n given points with a shape of the minimum area. One shape is parallelogram and the other is convex polygon.

1.4. Organization

In Section 2 we give the fundamental properties of convex layers and Tukey layers. In Section 3, we present the proof of the upper bound on the expected size of the first t Tukey layers, when the n points in X are sampled from a convex polygon. In Section 4, we prove the upper bound on the expected size of the first t convex layers under the same setting. In Section 5, we derive the lower bounds on the expected size of the first t Tukey layers for two special cases. Finally in Section 6, we apply our results to the average-case analysis of two shape fitting algorithms.

2. Preliminaries

In this section, we prepare some fundamental facts on Tukey Layers and convex layers. The readers are recommended to have a look through the statements to get familiar with these properties. Nonetheless, we include the proofs for completeness. To the best of our knowledge, the observations on Tukey layers are new and not found in the literature.

2.1. Convex layers, Tukey layers and their relation

The following lemma shows that the points in $U_t(X)$ are exactly the vertices of the t-th Tukey layer $S_t(X)$, which justifies referring the size of $S_t(X)$ to $|U_t(X)|$ as we mentioned after Definition 4.

Lemma 1. *For a planar point set X, the points in the t-th Tukey layer of X are in the position of a convex polygon. Equivalently, $U_t(X)$ has only one convex layer.*

Proof. Suppose there are at least two convex layers in $U_t(X)$. Let V_1 denote the vertices of the convex hull of $U_t(X)$, and $V_2 := U_t(X) \setminus V_1$. For any point $p \in V_2$, let ℓ be the line through p such that there are exactly $(t-1)$ points on one side. Notice that ℓ is through p and thus also through the interior of the convex hull of $U_1(X)$. Hence, on the side of ℓ that contains $(t-1)$ points, there must exist a point q which belongs to V_1. This implies that for the line ℓ' through q and parallel to ℓ, there are at most $(t-2)$ points on its one side. This contradicts the fact that $q \in V_1 \subseteq U_t(X)$. Finally we conclude that there can be only one single convex layer in each $U_t(X)$. □

The next lemma relates Tukey layers and convex layers.

Lemma 2. *It holds that $U_{[t]}(X) \subseteq V_{[t]}(X)$.*

Proof. If a point $p \in X \setminus V_{[t]}(X)$, then p can only lie on the $(t+1)$-st or a deeper layer of X. On any side of any line passing through p, there must be at least one vertex from each previous layer, including the 1-st to the t-th. In total there are at least t points and by Definition 4 it holds that $p \notin U_{[t]}(X)$. In conclusion, $U_{[t]}(X) \cap (X \setminus V_{[t]}(X)) = \emptyset$ and thus $U_{[t]}(X) \subseteq V_{[t]}(X)$. □

The following lemma discusses the relative position of Tukey layers. It shows that the vertices on the first t Tukey layers are outside the $(t+1)$-st Tukey layer.

Lemma 3. *It holds that $U_{[t]}(X) \cap S_{t+1}(X) = \emptyset$. As a consequence, $S_t(X) \subseteq H\left(X \setminus U_{[t-1]}(X)\right)$.*

Proof. Suppose not. We let $p \in U_{[t]}(X) \cap S_{t+1}(X)$ and ℓ be a line through p, on one side of which there are at most $(t-1)$ points.

If ℓ intersects the interior of $S_{t+1}(X)$, then there must be a $q \in U_{t+1}(X)$ on the side of ℓ where there are at most $(t-1)$ points. Let ℓ' denote the line through q and parallel to ℓ. Then there are at most $(t-2)$ points on one side of ℓ' and this contradicts the fact that $q \in U_{t+1}(X)$.

If ℓ does not intersect the interior of $S_{t+1}(X)$, then p must lie on a side rq of the boundary of $S_{t+1}(X)$. Here $r, q \in U_{t+1}(X)$ and the line segment rq must be on the line ℓ. As there are at most $(t-1)$ points on one side of ℓ, we then have $r, q \in U_{[t]}(X)$, contradictory to the assumption that rq is a side of the boundary of $S_{t+1}(X)$. □

Lemma 4. *If $X_1 \cup X_2 = X$, then $U_{[t]}(X) \subseteq U_{[t]}(X_1) \cup U_{[t]}(X_2)$.*

Proof. For each point $p \in U_{[t]}(X)$, there exists a line ℓ through it, on one side of which there are at most $(t-1)$ points of X. Then there will be neither more than $(t-1)$ points of X_1 nor more than $(t-1)$ points of X_2 on the same side of ℓ. Then we have $p \in U_{[t]}(X_1)$ when $p \in X_1$, and $p \in U_{[t]}(X_2)$ when $p \in X_2$. □

The following corollary is a generalization to k subsets.

Corollary 1. *Given $X = X_1 \cup X_2 \cup \cdots \cup X_k$, we have*

$$U_{[t]}(X) \subseteq U_{[t]}(X_1) \cup U_{[t]}(X_2) \cup \cdots \cup U_{[t]}(X_k).$$

The following lemma is an analogous result of Lemma 4 for $V_{[t]}$.

Lemma 5. *If $X_1 \cup X_2 = X$, then $H_t(X_1) \cup H_t(X_2) \subseteq H_t(X)$ and $V_{[t]}(X) \subseteq V_{[t]}(X_1) \cup V_{[t]}(X_2)$.*

Proof. We prove the lemma by induction on t. The statement is well-known when $t = 1$. Assume it holds for t and we shall prove it for $(t+1)$. By the induction hypothesis, $V_{[t]}(X) \subseteq V_{[t]}(X_1) \cup V_{[t]}(X \setminus X_1)$, we then have

$$X \setminus V_{[t]}(X) \supseteq X_1 \setminus V_{[t]}(X) \supseteq X_1 \setminus \left(V_{[t]}(X_1) \cup V_{[t]}(X \setminus X_1)\right) = X_1 \setminus V_{[t]}(X_1).$$

Further by Definition 1,

$$H_{t+1}(X_1) = H\left(X_1 \setminus V_{[t]}(X_1)\right) \subseteq H\left(X \setminus V_{[t]}(X)\right) = H_{t+1}(X).$$

Similarly, $H_{t+1}(X_2) \subseteq H_{t+1}(X)$. Therefore $H_{t+1}(X_1) \cup H_{t+1}(X_2) \subseteq H_{t+1}(X)$.

Now we prove $V_{[t+1]}(X) \subseteq V_{[t+1]}(X_1) \cup V_{[t+1]}(X_2)$. For a point $p \in V_{[t+1]}(X)$, p cannot be in the interior of $H_{t+1}(X)$. We have already shown that $H_{t+1}(X_1) \cup H_{t+1}(X_2) \subseteq H_{t+1}(X)$, so p cannot be in the interior of either $H_{t+1}(X_1)$ or $H_{t+1}(X_2)$. If $p \in X_1$, then $p \in V_{[t+1]}(X_1)$; otherwise $p \in V_{[t+1]}(X_2)$. □

Corollary 2. *Given $X = X_1 \cup X_2 \cup \cdots \cup X_k$, we have*

$$V_{[t]}(X) \subseteq V_{[t]}(X_1) \cup V_{[t]}(X_2) \cup \cdots \cup V_{[t]}(X_k).$$

2.2. Convex depth

The following lemma examines how the convex depth of a point p in X changes after an additional point q in added to X.

Lemma 6. *Given a planar point set X and a point $p \in X$, the convex depth of p will either remain unchanged or increase at most by 1 after an additional point q is added into X.*

Proof. By the proof of [8, Lemma 3.1], we know that $V_t(X) \subseteq V_t(X \cup \{q\}) \cup V_{t+1}(X \cup \{q\})$. For $p \in V_t(X)$, either $p \in V_t(X \cup \{q\})$ or $p \in V_{t+1}(X \cup \{q\})$. In other words, the convex depth of p will either remain unchanged or increase by 1. □

2.3. Expected area and expected size of Tukey layers

The following lemma shows the relation between the expected size and the expected area of the Tukey layers.

Lemma 7. *Let $C \subseteq \mathbb{R}^2$ be a bounded and closed convex set of unit area and X be the set of n points chosen independently and uniformly from C. Then*

$$\mathbb{E}\left|U_{[t]}(X)\right| \leq n\left[1 - \mathbb{E}\, A(S_{t+1}(X))\right].$$

Proof. On the one hand, by Lemma 3, the points in $U_{[t]}(X)$ must be outside $S_{t+1}(X)$. On the other hand, there might be points of $X \setminus U_{[t]}(X)$ not lying in $S_{t+1}(X)$, either. Since those points not belonging to $S_{t+1}(X)$ are uniform in $C \setminus S_{t+1}(X)$, in expectation we have

$$\mathbb{E}\left|U_{[t]}(X)\right| \leq n\,\mathbb{E}\left[1 - A(S_{t+1}(X))\right] = n\left[1 - \mathbb{E}\, A(S_{t+1}(X))\right]. \quad \square$$

2.4. Upper (lower) hull of Tukey layer

For a general convex polygon, let P_1 be the vertex with the smallest x-coordinate and Q_1 the vertex with the largest x-coordinate, where we break the tie by choosing the point with the largest y-coordinate for both points. Then, the *upper hull* refers to the boundary of the polygon from P_1 to Q_1 in the clockwise orientation. Similarly, let P_2 be the vertex with the smallest x-coordinate and Q_2 the vertex of the largest x-coordinate of the polygon, where we break the tie by choosing the point with the smallest y-coordinate. It may happen that $P_1 = P_2$ and $Q_1 = Q_2$. The *lower hull* refers to the boundary from Q_2 to P_2 in the clockwise orientation.

For a point P, if the ray ejecting vertically downwards (upwards) from P crosses the upper (lower) hull of the convex polygon, we shall say it is above (below) the upper (lower) hull.

3. Upper bound on expected size of Tukey layers

In this section, we prove $\mathbb{E}\left|U_{[t]}(X)\right| = O(kt\log(n/k))$, when the n points of X are sampled independently and uniformly from a convex k-gon. Our proof is inspired by [5] in which Har-Peled considered the expected size of the convex hull of X for X uniformly sampled from a triangle of unit area. He partitions the triangle into $n \times n$ equal-area cells and gives a lower bound on the expected number of cells that are inside the convex hull. Dividing the lower bound by n^2 would yield a lower

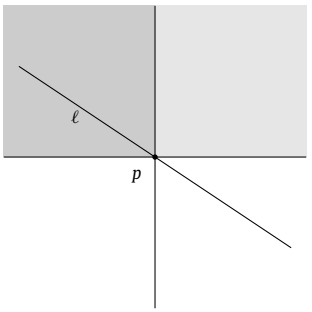

Fig. 2. The plane is divided into 4 open quadrants by the horizontal and vertical lines through p. The upper left and upper right quadrants are marked by dark gray and light gray color, respectively. Line ℓ is an arbitrary non-vertical line through p.

bound on the expected area of the convex hull, denoted by $\mathbb{E}\,A(X)$. Then by $\mathbb{E}\,|V(X)| = n\,[1 - \mathbb{E}\,A(X)]$ from [13], an upper bound on the expected size $\mathbb{E}\,|V(X)|$ of the convex hull follows. The case where X is uniformly sampled from a convex k-gon can be reduced to triangles by partitioning the k-gon into k triangles. Before proving our main results, we need the following auxiliary lemma.

Lemma 8. *Given a point $p \in X$, the plane is partitioned into four open quadrants by the horizontal and vertical lines through p, as shown in Fig. 2. If both the upper-left and upper-right quadrants contain at least t points of X, then for any non-vertical line ℓ through p, there must be at least t points of X above ℓ. In other words, the point p cannot be above the upper hull of $S_t(X)$.*

Proof. For any non-vertical line ℓ through p, either the upper-left or the upper-right quadrant is completely above ℓ. Since both quadrants contain at least t points, there are always t points above ℓ. By Definition 3, we know that p cannot be above the upper hull of the t-th Tukey layer. □

Since the points in X are chosen uniformly at random, we may assume that no three points are collinear and no two points have the same x or y coordinate, because such degenerate cases happen with zero probability. We decompose the convex hull into an upper hull and a lower hull, as defined in Section 2.4. Lemma 8 implies that

$$\begin{aligned}&\Pr(p \text{ is below the upper hull of } U_t(X))\\ &\qquad\geq \Pr(p \text{ has at least } t \text{ points in both upper-left and upper-right quadrants})\end{aligned}$$

and similarly

$$\begin{aligned}&\Pr(p \text{ is above the lower hull of } U_t(X))\\ &\qquad\geq \Pr(p \text{ has at least } t \text{ points in both lower-left and lower-right quadrants}).\end{aligned}$$

Then we can upper bound $\Pr\left(p \in X \setminus U_{[t]}(X)\right)$ as

$$\begin{aligned}&\Pr\left(p \in U_{[t]}(X)\right)\\ &= \Pr(p \text{ is on or above the upper hull of } U_t(X))\\ &\quad + \Pr(p \text{ is on or below the lower hull of } U_t(X))\\ &= (1 - \Pr(p \text{ is below the upper hull of } U_t(X)))\\ &\quad + (1 - \Pr(p \text{ is above the lower hull of } U_t(X))),\end{aligned}$$

whence an upper bound on $\Pr\left(p \in U_{[t]}(X)\right)$ would follow. Multiplying the upper bound by n would finally produce an upper bound on $\mathbb{E}\left|U_{[t]}(X)\right|$.

Theorem 1. *Let X be a set of n points sampled independently and uniformly from a triangle, then $\mathbb{E}\left|U_{[t-1]}(X)\right| \leq 4t \ln n + 4t + 10$.*

Denote the triangle by T and, without loss of generality, assume that T has area 1. We partition T into n equal-area triangles by segments emanating from a fixed vertex. Each triangle is further partitioned into one triangle and $(n-1)$ trapezoids with equal-area by line segments parallel to the opposite side. See Fig. 3 for an illustration. There are thus n^2 cells in T, each has area $1/n^2$. Let $G_{i,j}$ denote the cell in the i-th row and j-th column. We also define $G_{[i_1,i_2],[j_1,j_2]} = \bigcup_{i'=i_1}^{i_2} \bigcup_{j'=j_1}^{j_2} G_{i',j'}$.

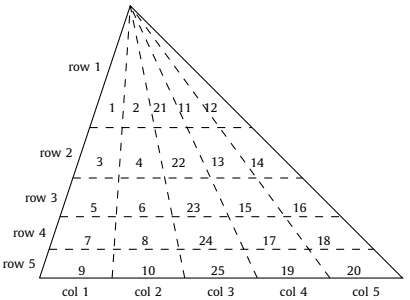

Fig. 3. Partitioning of a triangle into n^2 equal-area cells for $n = 5$. The cells are numbered for $j = 3$ by Eq. (1).

Proof of Theorem 1. We shall count the expected number of cells in each column that are above (resp. below) or intersecting the upper (resp. lower) hull of $S_t(X)$, the t-th Tukey layer. Summing up all those values will lead to an upper bound on the expected number of cells, and thus the expected area, outside $S_t(X)$. Since in an $n \times n$ grid, the boundary of a convex polygon can intersect at most $4n$ cells in total, we only need to count how many cells in the j-th column are above the upper hull of $S_t(X)$.

To count the expected number of cells above the upper hull of $S_t(X)$, let Z_j $(1 < j < n)$ denote the maximum i such that G_{ij} is above the upper hull of $S_t(X)$ and we shall find an upper bound on $\mathbb{E}[Z_j]$. Let I_1 (resp. I_2) be the row index of the t-th point from top to bottom in $G_{[1,n],[1,j-1]}$ (resp. $G_{[1,n],[j+1,n]}$). Then for any $G_{i,j}$ with $i > \max(I_1, I_2)$, there must be at least t points in its upper left quadrant and also t points in its upper right quadrant. By Lemma 8, such a point cannot be above the upper hull of $S_t(X)$. Therefore, $Z_j \le \max(I_1, I_2) \le I_1 + I_2$ and thus $\mathbb{E}\, Z_j \le \mathbb{E}\, I_1 + \mathbb{E}\, I_2$. We can prove that $\mathbb{E}\, I_1 \le \frac{tn}{j-1} + 1$ and $\mathbb{E}\, I_2 \le \frac{tn}{n-j} + 1$ (the proof is postponed to Lemma 9), then

$$\mathbb{E}\, Z_j \le \mathbb{E}\, I_1 + \mathbb{E}\, I_2 \le \frac{tn}{j-1} + \frac{tn}{n-j} + 2.$$

To count the expected number of cells below the lower hull of $S_t(X)$, we analogously define Z'_j to be the maximum i such that $G_{n-i+1,j}$ is below or intersects the lower hull of $S_t(X)$. A similar argument to the above shows the same upper bound on $\mathbb{E}[Z'_j]$, that is,

$$\mathbb{E}\, Z'_j \le \frac{tn}{j-1} + \frac{tn}{n-j} + 2.$$

Note that the first and the last column each contains at most n cells outside $S_t(X)$. The expected number of cells in T which are outside $S_t(X)$ is therefore at most

$$\begin{aligned} 2n + \sum_{j=2}^{n-1} \left(\mathbb{E}\, Z_j + \mathbb{E}\, Z'_j \right) &\le 2n + 2 \cdot \sum_{j=2}^{n-1} \left(\frac{tn}{j-1} + \frac{tn}{n-j} + 2 \right) \\ &\le 2n + 2\,[2tn \ln(n-2) + 2tn + 2(n-2)] \\ &\le 4tn \ln n + 4tn + 6n, \end{aligned}$$

together with the at most $4n$ cells that intersect the boundary of the t-th Tukey layer $S_t(X)$, when $n \ge 4$. It follows that

$$\mathbb{E}\, A(S_t(X)) \ge 1 - \frac{4tn \ln n + 4tn + 6n + 4n}{n^2} \ge 1 - \frac{4t \ln n + 4t + 10}{n}.$$

By Lemma 7, we finally conclude that $\mathbb{E}\left|U_{[t-1]}\right| \le 4t \ln n + 4t + 10$ when $n \ge 4$. When $n < 4$, this bound holds trivially since $\mathbb{E}\left|U_{[t-1]}\right| \le n$. □

Lemma 9. *Suppose that $1 < j < n$. Let I_1 (resp. I_2) be the row indices of the t-th point from top to bottom in $G_{[1,I_1],[1,j-1]}$ (resp. $G_{[1,I_2],[j+1,n]}$), then $\mathbb{E}\, I_1 \le \frac{tn}{j-1} + 1$ and $\mathbb{E}\, I_2 \le \frac{tn}{n-j} + 1$.*

Proof. We prove $\mathbb{E}\, I_1 \le \frac{tn}{j-1} + 1$ below, and a similar argument will give $\mathbb{E}\, I_2 \le \frac{tn}{n-j} + 1$. We number the n^2 cells from 1 to n^2 as follows. For a cell $G_{i,\ell}$, we define its number

$$idx(G_{i,\ell}) = \begin{cases} (j-1)(i-1) + \ell, & \ell < j; \\ (j-1)n + (n-j)(i-1) + \ell, & \ell > j; \\ (n-1)n + i, & \ell = j. \end{cases} \tag{1}$$

See Fig. 3 for an illustration. Intuitively, the triangle is split into three parts, left to the j-th column, right to the j-the column and the j-th column. In each part the cells are numbered one by one from left to right and from top to bottom; overall, the left part precedes the right part and the right part precedes the j-th column.

Now, we can refer to each cell by its number and denote the cells by $G_1, \ldots, G_{n^2}$, abusing the notation. Since all cells have the same area, a uniform random point in the triangle T can be generated by first choosing an integer in $m \in \{1, \ldots, n^2\}$ uniformly at random and then generating a uniform random point in G_m. Also we denote by $|G_m|$ the number of points in X that are contained in G_m.

Let h be the integer such that $\sum_{i=1}^{h-1} |G_i| < t$ and $\sum_{i=1}^{h} |G_i| \geq t$. This is exactly the t-th smallest integer among n uniform samples from $\{1, \ldots, n^2\}$. Let $f_t(x)$ be the density function of the t-th smallest value among n independent uniform points in $[0, 1]$. Then

$$\begin{aligned}\mathbb{E}\, h = \int_0^1 \lceil xn^2 \rceil f_t(x)dx \leq \int_0^1 (xn^2+1) f_t(x)dx &= n^2 \int_0^1 x f_t(x)dx + 1 \\ &= n^2 \frac{t}{n+1} + 1 \\ &\leq t(n-1) + 1 \\ &\leq tn.\end{aligned}$$

Here we used the fact that $\int_0^1 x f_t(x)dx = \frac{t}{n+1}$. The integral is the expected value of the t-th smallest value among n independent uniform points in $[0, 1]$, and it is a classic result that this expected value is exactly $t/(n+1)$ (see, e.g., [23, Lemma 8.3]).

When $h \leq n(j-1)$, we have $I_1 = \lceil h/(j-1) \rceil$. When $h > n(j-1)$, it automatically holds that $I_1 \leq n \leq h/(j-1)$. In both cases, we have $I_1 \leq \lceil h/(j-1) \rceil$. Therefore,

$$\mathbb{E}\, I_1 \leq \mathbb{E}\left\lceil \frac{h}{j-1} \right\rceil \leq \frac{\mathbb{E}\, h}{j-1} + 1 \leq \frac{tn}{j-1} + 1. \quad \square$$

Theorem 2. *Let X be a set of n points sampled independently and uniformly from a convex k-gon. Then we have $\mathbb{E}\left|U_{\lfloor t-1 \rfloor}(X)\right| \leq 4tk \ln(n/k) + 4tk + 10k$.*

Proof. Partition the convex k-gon into k triangles. Let $X_1, X_2, \ldots, X_k$ be the set of points of X in the triangles and $n_i = |X_i|$ for $i = 1, \ldots, k$. Note that $n_1, n_2, \ldots, n_k$ are random numbers subject to $\sum_{i=1}^{k} n_i = n$. It follows from Corollary 1 that

$$\begin{aligned}\mathbb{E}\left[U_{\lfloor t-1 \rfloor}(X) | n_1, n_2, \ldots, n_k\right] \leq \sum_{i=1}^{k} \mathbb{E}\left[U_{\lfloor t-1 \rfloor}(X_i) | n_i\right] &\leq \sum_{i=1}^{k} (4t \ln n_i + 4t + 10) \\ &= 4t \sum_{i=1}^{k} \ln n_i + 4tk + 10k \\ &\leq 4tk \ln(n/k) + 4tk + 10k. \quad \square\end{aligned}$$

4. Upper bound on expected size of convex layers

In this section, we shall prove an upper bound $O\left(kt^3 \log \frac{n}{kt^2}\right)$ on $\mathbb{E}\left|V_{[t]}(X)\right|$, when X is sampled uniformly from a convex k-gon. The proof is inspired by [7] and [21]. We first consider the case where the points in X are sampled uniformly from a triangle T and obtain an upper bound $O\left(kt^3 \log \frac{n}{kt^2}\right)$, which, by Corollary 2, implies an upper bound $O\left(kt^3 \log \frac{n}{kt^2}\right)$ when X is sampled from a k-gon. The problem can be further reduced to finding an upper bound on the probability $\Pr\left(p \in V_{[t]}(X)\right)$ for a single point $p \in X$, which, multiplied by n, will be an upper bound on $\mathbb{E}\left|V_{[t]}(X)\right|$.

Theorem 3. *Let X be a set of n points sampled independently and uniformly from a triangle T, then $\mathbb{E}\left|V_{[t]}(X)\right| = O\left(t^3 \log(n/t^2)\right)$.*

Proof. As the combinatorial properties of convex hulls are affine invariant, we may assume the vertices of T are $(0, 0)$, $(1, 0)$ and $(0, 1)$. We partition T into three regions R_1, R_2, R_3 with equal area by connecting the centroid $\left(\frac{1}{3}, \frac{1}{3}\right)$ to the midpoint of each edge (see Fig. 4). Then $\Pr\left(p \in V_{[t]}(X) | p \in R_i\right)$ are all equal for $i = 1, 2, 3$ and so

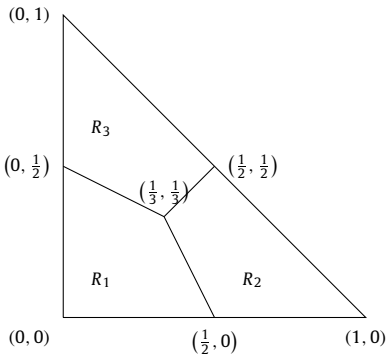

Fig. 4. The triangle is divided into three parts, by connecting the centroid to the midpoint of each edge.

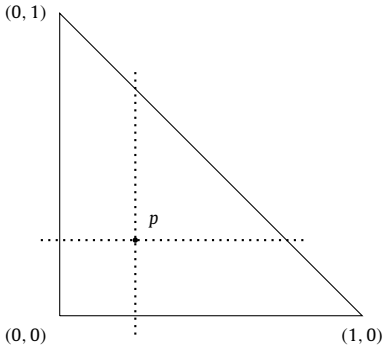

Fig. 5. By the horizontal line and the vertical line through a given point p, the triangle is divided into four quadrants.

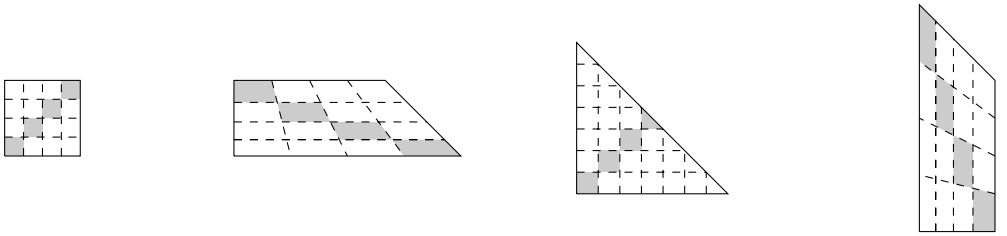

Fig. 6. Partition of each quadrant of the triangle into cells when $t = 4$. In each single quadrant, the cells have the equal area. There are exactly t diagonal cells in each quadrant, marked in gray color.

$$\begin{aligned}\Pr\left(p \in V_{[t]}(X)\right) &= \sum_{i=1}^{3} \Pr\left(p \in V_{[t]}(X) | p \in R_i\right) \Pr(p \in R_i) \\ &= \sum_{i=1}^{3} \Pr\left(p \in V_{[t]}(X) | p \in R_i\right) \cdot \frac{1}{3} \\ &= \Pr\left(p \in V_{[t]}(X) | p \in R_1\right).\end{aligned}$$

We turn to find an upper bound on $\Pr\left(p \in V_{[t]}(X) | p \in R_1\right)$. For this purpose, the triangle T is divided into four quadrants by a vertical and a horizontal line through p as shown in Fig. 5. Each quadrant is further partitioned into multiple cells as in Fig. 6. The triangular quadrant is partitioned into $(2t+1)t$ cells by $(2t-1)$ equally spaced horizontal lines and another $(2t-1)$ equally spaced vertical lines. Each of the other three quadrants are partitioned into t^2 equal-area cells. This construction ensures exactly t diagonal cells in each of the four quadrants.

We claim that if $p \in V_{[t]}(X)$, then at least one of the $4t$ diagonal cells must be empty. The proof of this claim is deferred to Lemma 10. By this observation, the probability of $p \in V_{[t]}(X)$ is at most the probability that at least one of the $4t$ diagonal cells is empty, which we upper bound as follows. Let (p_1, p_2) denote the coordinates of p. When $p \in R_1$, the area of each quadrant is at least $\frac{1}{2}p_1p_2$ by [7, Section 2] and the probability mass (with respect to the uniform distribution on T) of each quadrant is at least p_1p_2. Therefore each diagonal cell has probability mass at least $\frac{p_1p_2}{4t^2}$, and the expected number of

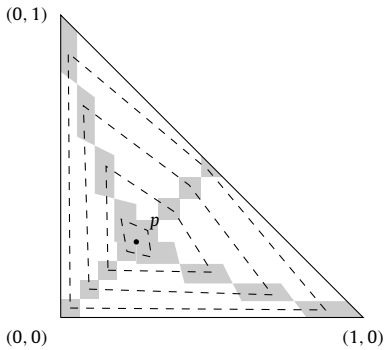

Fig. 7. The diagonal cells are shaded. Connecting one point in the diagonal cell of the same order in each quadrant forms a convex layer, marked by a dashed polyline.

points in every single cell is at least $\frac{np_1p_2}{4t^2}$. By the multiplicative form of Chernoff bound [23, Theorem 4.5], the probability that a diagonal cell is empty is at most $\exp\left(-\frac{np_1p_2}{16t^2}\right)$. Further by a union bound, the probability that at least one of the $4t$ diagonal cells is empty in triangle T is at most $4t\exp\left(-\frac{np_1p_2}{16t^2}\right)$. Therefore,

$$\Pr\left(p \in V_{[t]}(X) | p_1p_2 = y, p \in R_1\right) \leq 4te^{-\frac{ny}{16t^2}},$$

whence we can show that

$$\Pr\left(p \in V_{[t]}(X) | p \in R_1\right) \leq 12t \int_0^{1/9} e^{-\frac{ny}{16t^2}} \log\frac{1}{y}\,\mathrm{d}y = 12t \cdot O\left(\frac{t^2}{n}\log\frac{n}{t^2}\right),$$

whose proof is postponed to Lemma 13 and Lemma 14. It follows that $\Pr\left(p \in V_{[t]}(X)\right) = O\left(\frac{t^3}{n}\log\frac{n}{t^2}\right)$ for any $p \in X$ and, finally, that $\mathbb{E}\left|V_{[t]}(X)\right| = O\left(t^3\log\frac{n}{t^2}\right)$. □

Now we are ready to prove the following main theorem.

Theorem 4. *Let X be a set of n points sampled independently and uniformly from a convex k-gon, then we have $\mathbb{E}\left|V_{[t]}(X)\right| = O\left(kt^3\log\frac{n}{kt^2}\right)$.*

Proof. As in the proof of Theorem 2, we partition the k-gon into k triangles. Let $n_1, n_2, \ldots, n_k$ denote the number of points in each triangle. It follows from Corollary 2 that

$$\begin{aligned}\mathbb{E}\left[V_{[t]}(X) | n_1, n_2, \ldots, n_k\right] &\leq \sum_{i=1}^{k} \mathbb{E}\left[V_{[t]}(X_i) | n_i\right] \leq \sum_{i=1}^{k} O\left(t^3\log\frac{n_i}{t^2}\right) \\ &= O\left(kt^3\log\frac{n}{kt^2}\right),\end{aligned}$$

where we used the AM-GM inequality and the fact that $\sum_{i=1}^{k} n_i = n$ in the last step. □

In the rest of this section, we state and prove those lemmata used in the proof of Theorem 3. We denote the density and the cumulative distribution functions of the product $p_1 \cdot p_2$ by $\rho_{p_1p_2}(\cdot)$ and $F_{p_1p_2}(\cdot)$, respectively.

Lemma 10. *If $p \in V_{[t]}(X)$, there must be at least one empty diagonal cell.*

Proof. If none of the $4t$ diagonal cells is empty, we can construct t convex layers enclosing p, where each layer consists of four points from the diagonal cells, one from each quadrant (see Fig. 7). The convex depth of p is thus at least $(t+1)$. Although there may be more than one point in each diagonal cell, we know from Lemma 6 that the convex depth of p cannot decrease after those additional points are included. This contradicts the assumption that $p \in V_{[t]}(X)$. Therefore, some diagonal cell must be empty. □

Lemma 11 *([7, Theorem 1]).* $F_{p_1p_2}(y|p \in R_1) \le 3F_{p_1p_2}(y|p \in [0,1]\times[0,1])$.

Lemma 12 *([24, section I.8]).* $\rho_{p_1p_2}(y|p \in [0,1]\times[0,1]) = \log(1/y)$.

Lemma 13. *If* $\Pr\left(p \in V_{[t]}(X)|p_1p_2 = y, p \in R_1\right) \le 4te^{-\frac{ny}{16t^2}}$, *then*

$$\Pr\left(p \in V_{[t]}(X)|p \in R_1\right) \le 12t \int_0^{1/9} e^{-\frac{ny}{16t^2}} \log\frac{1}{y}\,\mathrm{d}y.$$

Proof. It is easy to prove that p_1p_2 reaches its maximum value $\frac{1}{9}$ at $\left(\frac{1}{3},\frac{1}{3}\right)$ for $p \in H_1$. Then we have

$$\begin{aligned}\Pr\left(p \in V_{[t]}(X)|p \in R_1\right) &= \int_0^{1/9} \Pr\left(p \in V_{[t]}(X)|p_1p_2 = y, p \in R_1\right)\cdot \rho_{p_1p_2}(y|p \in R_1)\,\mathrm{d}y\\ &\le 4t\int_0^{1/9} e^{-\frac{ny}{16t^2}}\cdot \rho_{p_1p_2}(y|p \in R_1)\,\mathrm{d}y\\ &= 4t\int_0^{1/9} e^{-\frac{ny}{16t^2}}\,\mathrm{d}F_{p_1p_2}(y|p \in R_1).\end{aligned}$$

By Lemma 11 and Lemma 12,

$$\begin{aligned}\int_0^{1/9} e^{-\frac{ny}{16t^2}}\,\mathrm{d}F_{p_1p_2}(y|p \in R_1) &\le 3\int_0^{1/9} e^{-\frac{ny}{16t^2}}\,\mathrm{d}F_{p_1p_2}(y|p \in [0,1]\times[0,1])\\ &= 3\int_0^{1/9} e^{-\frac{ny}{16t^2}}\log\frac{1}{y}\,\mathrm{d}y.\end{aligned}$$

Thus

$$\Pr\left(p \in V_{[t]}(X)|p \in R_1\right) \le 12t\int_0^{1/9} e^{-\frac{ny}{16t^2}}\log\frac{1}{y}\,\mathrm{d}y. \quad \square$$

Lemma 14. $\int_0^{1/9} e^{-\frac{ny}{16t^2}}\log\frac{1}{y}\,\mathrm{d}y = O\left(\frac{t^2}{n}\log\frac{n}{t^2}\right)$.

Proof. Substituting y with $z = \frac{ny}{16t^2}$, we have

$$\begin{aligned}I &= \int_0^{1/9} e^{-\frac{ny}{16t^2}}\log\frac{1}{y}\,\mathrm{d}y\\ &= \frac{16t^2}{n}\int_0^{1/9} e^{-z}\left(\log\frac{n}{16t^2}+\log\frac{1}{z}\right)\mathrm{d}z\\ &\le \frac{16t^2}{n}\log\frac{n}{16t^2}\int_0^{\infty} e^{-z}\,\mathrm{d}z + \frac{16t^2}{n}\int_0^{\infty} e^{-z}\log\frac{1}{z}\,\mathrm{d}z.\end{aligned}$$

Since both $\int_0^{\infty} e^{-z}dz$ and $\int_0^{\infty} e^{-z}\log\frac{1}{z}\,\mathrm{d}z$ are constants, we conclude

$$\int_0^{1/9} e^{-\frac{ny}{16t^2}}\log\frac{1}{y}\,\mathrm{d}y = O\left(\frac{t^2}{n}\log\frac{n}{t^2}\right). \quad \square$$

5. Lower bound of expected size of Tukey layers

We shall prove the lower bound on the expected size of $U_{[t]}(X)$, the first t Tukey layers, for two special cases where X is sampled from a parallelogram (Section 5.1) and a triangle (Section 5.2). We need the following lemma throughout this section.

Lemma 15 *([6, Section 3]). For all integer $r, s \geq 0$ and for all $c \in (0, 1]$ we have*

$$\int_0^1 \int_0^1 (1 - cxy)^{n-s} (xy)^r \mathrm{d}x\mathrm{d}y = \frac{r!}{c^{r+1}} \cdot \frac{\log n}{n^{r+1}} + O\left(\frac{1}{n^{r+1}}\right), \quad n \to \infty.$$

5.1. Parallelogram

Without loss of generality, we may assume that the parallelogram is a unit square $[0, 1] \times [0, 1]$, because the combinatorial properties would not change under an affine transformation. For each point $p = (p_1, p_2) \in X$, we now compute the probability that it is on the first t Tukey layers of X. For this purpose, we introduce the following definition.

Definition 5. Given a point $p = (p_1, p_2)$ with $0 \leq p_1 < \frac{1}{2}$ and $0 \leq p_2 < \frac{1}{2}$, the dividing line is defined to be

$$\ell_0 : \frac{x}{2p_1} + \frac{y}{2p_2} = 1.$$

The dividing line when $p_1 \geq \frac{1}{2}$ or $p_2 \geq \frac{1}{2}$ can be defined symmetrically.

The line divides the unit square into a triangle of area $2p_1p_2$ and a pentagon of area $(1 - 2p_1p_2)$. Notice that a sufficient condition for a point p to be on the first t Tukey layers is that, there are no more than $(t - 1)$ points in the triangular part. We thus have the following theorem.

Theorem 5. *Suppose that X consists of n independent and uniformly sampled points from a unit square. There exists an absolute constant $\alpha > 0$ such that whenever $t \leq \alpha\sqrt{n}$, it holds that $\mathbb{E}\left|U_{[t]}(X)\right| = \Omega(t \log n)$ as $n \to \infty$. Furthermore, when $t = o((n/\log n)^{1/3})$, it holds that $\mathbb{E}\left|U_{[t]}(X)\right| \geq 2t \log n + O(1)$ as $n \to \infty$.*

Proof.

$$\begin{aligned}
\Pr(p \in U_{[t]}) &\geq \Pr(\text{no more than } t \text{ points under the dividing line } \ell_0) \\
&= 4\int_0^{\frac{1}{2}} \int_0^{\frac{1}{2}} \sum_{i=0}^{t-1} \binom{n-1}{i} (2p_1p_2)^i (1 - 2p_1p_2)^{n-1-i} \,\mathrm{d}p_1\,\mathrm{d}p_2 \\
&= 4\sum_{i=0}^{t-1} \binom{n-1}{i} \int_0^{\frac{1}{2}} \int_0^{\frac{1}{2}} (2p_1p_2)^i (1 - 2p_1p_2)^{n-1-i} \,\mathrm{d}p_1\,\mathrm{d}p_2 \\
&= \sum_{i=0}^{t-1} \binom{n-1}{i} \int_0^{\frac{1}{2}} \int_0^{\frac{1}{2}} (2p_1p_2)^i (1 - 2p_1p_2)^{n-1-i} \,\mathrm{d}(2p_1)\,\mathrm{d}(2p_2) \\
&= \sum_{i=0}^{t-1} \binom{n-1}{i} \int_0^1 \int_0^1 \left(\frac{p_1p_2}{2}\right)^i \left(1 - \frac{p_1p_2}{2}\right)^{n-1-i} \,\mathrm{d}p_1\,\mathrm{d}p_2 \\
&= \sum_{i=0}^{t-1} \frac{1}{2^i} \binom{n-1}{i} \int_0^1 \int_0^1 (p_1p_2)^i \left(1 - \frac{1}{2}p_1p_2\right)^{n-1-i} \,\mathrm{d}p_1\,\mathrm{d}p_2.
\end{aligned}$$

By Lemma 15, when $n \to \infty$, we have

$$\int_0^1 \int_0^1 (p_1p_2)^i \left(1 - \frac{1}{2}p_1p_2\right)^{n-1-i} \,\mathrm{d}p_1\,\mathrm{d}p_2 = \frac{i!}{\left(\frac{1}{2}\right)^{i+1}} \frac{\log n}{n^{i+1}} + O\left(\frac{1}{n^{i+1}}\right).$$

Therefore, as $n \to \infty$,

$$\begin{aligned}
\Pr(p \in U_{[t]}) &\geq \sum_{i=0}^{t-1} \frac{1}{2^i} \binom{n-1}{i} \left[\frac{i!}{\left(\frac{1}{2}\right)^{i+1}} \frac{\log n}{n^{i+1}} + O\left(\frac{1}{n^{i+1}}\right) \right] \\
&= \sum_{i=0}^{t-1} \left[2 \cdot \frac{(n-1)!}{(n-1-i)! \cdot n^i} \cdot \frac{\log n}{n} + O\left(\frac{1}{2^i i! n}\right) \right] \\
&= \sum_{i=0}^{t-1} \left[\frac{2\log n}{n} \cdot \left(1 - \frac{1}{n}\right) \cdot \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{i}{n}\right) + O\left(\frac{1}{2^i i! n}\right) \right] \\
&\geq \sum_{i=0}^{t-1} \frac{2\log n}{n} \cdot \left(1 - \frac{(i+1)i}{2n}\right) + O\left(\frac{1}{n}\right) \\
&\geq \sum_{i=0}^{t-1} \frac{2\log n}{n} \cdot \left(1 - \frac{(t-1)t}{2n}\right) + O\left(\frac{1}{n}\right) \\
&\geq \frac{2t\log n}{n} \left(1 - \frac{t^2}{2n}\right) + O\left(\frac{1}{n}\right).
\end{aligned}$$

Finally, the expected number of points on the first t Tukey layers

$$\mathbb{E}\left|U_{[t]}\right| = \sum_{p \in X} \Pr(p \in U_{[t]}) \geq 2\left(1 - \frac{t^2}{2n}\right) t\log n + O(1).$$

The conclusions follow immediately. □

5.2. Triangle

Theorem 6. *Suppose that X consists of n independent and uniformly sampled points from a triangle. There exists an absolute constant $\alpha > 0$ such that whenever $t \leq \alpha\sqrt{n}$, it holds that $\mathbb{E}\left|U_{[t]}(X)\right| = \Omega(t\log n)$ as $n \to \infty$.*

Proof. The proof is similar to that of Theorem 5. Without loss of generality, we assume that the vertices of the triangle are $(0,0)$, $(0,1)$ and $(1,0)$. Here we only consider those p where $0 \leq p_1 \leq \frac{1}{2}$ and $0 \leq p_2 \leq \frac{1}{2}$. We now find a lower bound on $\Pr\left(p \in U_{[t]}, 0 \leq p_1 \leq \frac{1}{2}, 0 \leq p_2 \leq \frac{1}{2}\right)$. Note that dividing line divides the triangle into a triangle of area $2p_1p_2$ and a quadrilateral of area $\frac{1}{2} - 2p_1p_2$. Their probability masses are $4p_1p_2$ and $1 - 4p_1p_2$ respectively.

$$\begin{aligned}
\Pr(p \in U_{[t]}) &\geq \Pr\left(p \in U_{[t]}, 0 \leq p_1 \leq \frac{1}{2}, 0 \leq p_2 \leq \frac{1}{2}\right) \\
&\geq \Pr(\text{no more than } t \text{ points under the dividing line } \ell_0) \\
&= \int_0^{\frac{1}{2}} \int_0^{\frac{1}{2}} \sum_{i=0}^{t-1} \binom{n-1}{i} (4p_1p_2)^i (1 - 4p_1p_2)^{n-1-i} \,\mathrm{d}p_1 \,\mathrm{d}p_2 \\
&= \frac{1}{4} \cdot \int_0^{\frac{1}{2}} \int_0^{\frac{1}{2}} \sum_{i=0}^{t-1} \binom{n-1}{i} (2p_1 \cdot 2p_2)^i (1 - 2p_1 \cdot 2p_2)^{n-1-i} \,\mathrm{d}(2p_1) \,\mathrm{d}(2p_2) \\
&= \frac{1}{4} \cdot \sum_{i=0}^{t-1} \binom{n-1}{i} \int_0^1 \int_0^1 (p_1p_2)^i (1 - p_1p_2)^{n-1-i} \,\mathrm{d}p_1 \,\mathrm{d}p_2
\end{aligned}$$

By Lemma 15, as $n \to \infty$,

$$\int_0^1 \int_0^1 (p_1p_2)^i (1 - p_1p_2)^{n-1-i} \,\mathrm{d}p_1 \,\mathrm{d}p_2 = \frac{i!\log n}{n^{i+1}} + O\left(\frac{1}{n^{i+1}}\right).$$

Therefore

$$\begin{aligned}
\Pr(p \in U_{[t]}) &\geq \frac{1}{4} \cdot \sum_{i=0}^{t-1} \left[\frac{(n-1)!}{n^i (n-i-1)!} \frac{\log n}{n} + O\left(\frac{1}{i!n}\right) \right] \\
&= \frac{1}{4} \cdot \sum_{i=0}^{t-1} \left[\left(1-\frac{1}{n}\right) \cdot \left(1-\frac{2}{n}\right) \cdots \left(1-\frac{i}{n}\right) \cdot \frac{\log n}{n} \cdot + O\left(\frac{1}{i!n}\right) \right] \\
&\geq \frac{1}{4} \cdot \sum_{i=0}^{t-1} \left[\left(1-\frac{(i+1)i}{2n}\right) \cdot \frac{\log n}{n} + O\left(\frac{1}{i!n}\right) \right] \\
&\geq \frac{1}{4} \cdot \sum_{i=0}^{t-1} \left[\left(1-\frac{(t-1)t}{2n}\right) \cdot \frac{\log n}{n} + O\left(\frac{1}{i!n}\right) \right] \\
&\geq \frac{1}{4} \cdot \frac{t \log n}{n} \cdot \left(1-\frac{t^2}{2n}\right) + O\left(\frac{1}{n}\right).
\end{aligned}$$

Finally, the expected number of points on the first t Tukey layers

$$\mathbb{E}\left|U_{[t]}\right| = \sum_{p \in X} \Pr(p \in U_{[t]}) \geq \frac{1}{4}\left(1-\frac{t^2}{2n}\right) t \log n + O(1).$$

The conclusions follow immediately. □

6. Applications

In this section, we discuss how our results in Sections 3 and 4 help in the average case analysis of two partial enclosing problems. The objective is to enclose $(n-t)$ of the given n points in X by a specified shape such that the area of the shape is minimized. This kind of problem is known as partial shape fitting and is an important problem in computational geometry, see, e.g., [9,25,26,12,27]. The points that are not enclosed are referred to as outliers [9,26,12].

The average case complexity is another important measure in addition to the worst case complexity. As pointed out in [3], the average case analysis is desirable because the best-case and worst-case performance of an algorithm usually differs greatly, especially for output-sensitive algorithms. In such situation, the average case complexity seems to be a more accurate and fair measurement of an algorithm's performance. A common scenario is that the input point set is drawn from some probability distribution and it is widely adopted by the computational geometry community to consider the uniform distribution in a convex polygon [6,14,15,7,13,5].

6.1. Enclosing parallelogram with minimum area

The algorithm given in [16] studies how to find a parallelogram with the minimum-area that encloses $(n-t)$ of the n given points. The time complexity of the algorithm is $O\left(t^4\tau^2 + n^2 \log n\right)$, where τ is the number of points whose Tukey depth is at most $(t+1)$. Such points coincide with $U_{[t+1]}(X)$ and so $\tau = \left|U_{[t+1]}(X)\right|$. In the worst case, $\left|U_{[t+1]}(X)\right| = n$ can be true and the worst case time complexity is thus $O\left(n^2 t^4 + n^2 \log n\right)$. However, on average, we have

$$\begin{aligned}
\mathbb{E}\left[O\left(t^4 \left|U_{[t]}(X)\right|^2 + n^2 \log n\right)\right] &\leq \mathbb{E}\left[O\left(nt^4 \left|U_{[t]}(X)\right| + n^2 \log n\right)\right] \\
&= O\left(kt^5 n \log \frac{n}{k} + n^2 \log n\right),
\end{aligned}$$

when X is uniformly sampled from a k-gon. When t is between $\Omega\left(\log^{\frac{1}{4}} n\right)$ and $O\left(\frac{n}{k \log \frac{n}{k}}\right)$, the average case complexity is smaller than the worst-case complexity. This explains why in many cases the actual runtime of the algorithm is faster than the worst-case complexity.

6.2. Minimum enclosing convex hull

Another application of our result is the algorithm for the minimum enclosing convex hull. Let X be a set of n points in $\mathbb{R}^2$. The problem asks to find a subset $X' \subset X$, $|X'| = t$, such that area of $H_t(X \setminus X')$ is minimized. In [9], Atanassov et al. provide an elegant solution to this problem with running time $O\left(n \log n + \binom{4t}{2t}(3t)^t |H_{[t]}(X)|\right)$. In the worst case, $|H_{[t]}(X)| = n$, which happens when X has at most t layers. For the average case, Theorem 4 implies a time complexity of $O\left(n \log n + k\binom{4t}{2t}(3t)^t t^3 \log \frac{n}{kt^2}\right)$, when X is uniformly distributed in convex k-gon. The average case is substantially better than the worst case when $t = O\left(\left(\frac{n}{k \log(n/k)}\right)^{1/3}\right)$.

7. Closing remarks

In this paper, we studied the expected size of the random convex layers and random Tukey layers of a point set X consisting of n points drawn independently and uniformly from a convex k-gon.

For random Tukey layers, we showed that $\mathbb{E}\,|U_{[t]}(X)| = O\,(kt\log(n/k))$ but only showed a matching lower bound of $\Omega(t\log n)$ for triangles and parallelograms. We leave an open problem of obtaining a general lower bound of $\Omega(kt\log n)$, for which a straightforward extension of our current technique of considering a line passing through a single point p in Section 5 seems inadequate. We also leave an open problem of obtaining a tight constant in the asymptotic results (which could depend on t); our constants are 4 in the upper bound and 2 in the lower bound, which are not tight since the tight constant is known to be $8/3$ when $t = 1$ [2].

For random convex layers, we showed that $\mathbb{E}\,|V_{[t]}(X)| = O\,(kt^3\log(n/(kt^2)))$. However, when the points are from sampled from a square, a better upper bound of $O\,(t^2\log(n/t^2))$ is known [21]. Thus, a natural question is whether it holds $\mathbb{E}\,|V_{[t]}(X)| = O\,(kt^2\log(n/(kt^2)))$ in general. Another interesting open problem is to obtain a lower bound with dependence on t, as existing lower bounds are only for $t = 1$ and there seem substantial difficulties to extend the existing techniques to a larger t.

CRediT authorship contribution statement

Zhengyang Guo: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Yi Li:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Shaoyu Pei:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J.W. Tukey, Mathematics and the picturing of data, in: Proceedings of the International Congress of Mathematicians, Vol. 2, Vancouver, 1975, 1975, pp. 523–531.
[2] A. Rényi, R. Sulanke, Über die konvexe hülle von n zufällig gewählten, Punkten, Z. Wahrscheinlichkeitstheor. Verw. Geb. 2 (1963) 75–84.
[3] R.A. Dwyer, Average-Case Analysis of Algorithms for Convex Hulls and Voronoi Diagrams, Citeseer, 1988.
[4] I. Hueter, The convex hull of a normal sample, Adv. Appl. Probab. 26 (4) (1994) 855–875.
[5] S. Har-Peled, On the expected complexity of random convex hulls, arXiv:1111.5340 [cs.CG], 2011.
[6] F. Affentranger, J.A. Wieacker, On the convex hull of uniform random points in a simpled-polytope, Discrete Comput. Geom. 6 (3) (1991) 291–305.
[7] R.A. Dwyer, On the convex hull of random points in a polytope, J. Appl. Probab. 25 (4) (1988) 688–699.
[8] K. Dalal, Counting the onion, Random Struct. Algorithms 24 (2) (2004) 155–165.
[9] R. Atanassov, P. Bose, M. Couture, A. Maheshwari, P. Morin, M. Paquette, M. Smid, S. Wuhrer, Algorithms for optimal outlier removal, J. Discret. Algorithms 7 (2) (2009) 239–248.
[10] S. Guha, Y. Li, Q. Zhang, Distributed partial clustering, ACM Trans. Parallel Comput. (TOPC) 6 (3) (2019) 1–20.
[11] S. Gupta, R. Kumar, K. Lu, B. Moseley, S. Vassilvitskii, Local search methods for k-means with outliers, Proc. VLDB Endow. 10 (7) (2017) 757–768.
[12] S. Har-Peled, Y. Wang, Shape fitting with outliers, SIAM J. Comput. 33 (2) (2004) 269–285.
[13] B. Efron, The convex hull of a random set of points, Biometrika 52 (3–4) (1965) 331–343.
[14] I. Bárány, et al., Sylvester's question: the probability that n points are in convex position, Ann. Probab. 27 (4) (1999) 2020–2034.
[15] C. Buchta, On the boundary structure of the convex hull of random points, Adv. Geom. 12 (1) (2012) 79–190.
[16] Z. Guo, Y. Li, Minimum enclosing parallelogram with outliers, arXiv:2003.01900 [cs.CG], 2020.
[17] B. Chazelle, On the convex layers of a planar set, IEEE Trans. Inf. Theory 31 (4) (1985) 509–517.
[18] I. Bárány, D.G. Larman, Convex bodies, economic cap coverings, random polytopes, Mathematika 35 (2) (1988) 274–291.
[19] B. Massé, On the LLN for the number of vertices of a random convex hull, Adv. Appl. Probab. 32 (3) (2000) 675–681.
[20] C. Schütt, The convex floating body and polyhedral approximation, Isr. J. Math. 73 (1) (1991) 65–77.
[21] M. He, C.P. Nguyen, N. Zeh, Maximal and convex layers of random point sets, in: Latin American Symposium on Theoretical Informatics, Springer, 2018, pp. 597–610.
[22] H. Raynaud, Sur l'enveloppe convexe des nuages de points aléatoires dans $\mathbf{R}^n$. I, J. Appl. Probab. 7 (1) (1970) 35–48.
[23] M. Mitzenmacher, E. Upfal, Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis, 2nd edition, Cambridge University Press, 2017.
[24] W. Feller, An Introduction to Probability Theory and Its Applications, Vol. 2, 2nd edition, 1971.
[25] S. Das, P.P. Goswami, S.C. Nandy, Smallest k-point enclosing rectangle and square of arbitrary orientation, Inf. Process. Lett. 94 (6) (2005) 259–266.
[26] H. Ding, A sub-linear time framework for geometric optimization with outliers in high dimensions, in: F. Grandoni, G. Herman, P. Sanders (Eds.), 28th Annual European Symposium on Algorithms (ESA 2020), in: Leibniz International Proceedings in Informatics (LIPIcs), vol. 173, Schloss Dagstuhl–Leibniz-Zentrumfür Informatik, Dagstuhl, Germany, 2020, 38.
[27] M. Segal, K. Kedem, Enclosing k points in the smallest axis parallel rectangle, Inf. Process. Lett. 65 (2) (1998) 95–99.