Computing the Expected Area of an Induced Triangle

Vissarion Fisikopoulos* Frank Staals§ Constantinos Tsirogiannis§

1 Introduction

Consider the following problem: given a set P of n points in the plane, compute the expected area of a triangle *induced by* P, that is, a triangle whose vertices are selected uniformly at random from the points in P. This problem is a special case of computing the expected area of the convex hull of k points, selected uniformly at random from P. These problems are important in computing the *functional diversity* in Ecology [4]. In this setting, each point represents some characteristics of a species, and the expected area of the convex hull provides an estimate of the diversity of the species, given that only k species exist in a geographic region.

We present a simple exact algorithm for the problem that computes the expected triangle area in $O(n^2 \log n)$ time, and extends to the case of computing the area of the convex hull of a size k subset. Additionally, we present a $(1 \pm \varepsilon)$-approximation algorithm for the case in which the ratio ρ between the furthest pair distance and the closest pair distance of the points in P is bounded. With high probability (whp.) our algorithm computes an answer in the range $[(1-\varepsilon)A^*, (1+\varepsilon)A]$, where A is the true expected triangle area, in $O(\frac{1}{\varepsilon^{8/3}} \rho^4 n^{5/3} \log^{4/3} n)$ expected time.

Notation. Let Δ denote the random variable corresponding to a triangle induced by P, and let $\mathcal{A}(Q)$ denote the area of a region $Q \subset \mathbb{R}^2$. We are thus interested in computing $\mathbb{E}[\mathcal{A}(\Delta)]$. We denote the probability of an event X by $\mathbb{P}[X]$. Assume w.l.o.g. that the origin o lies outside of the convex hull $\mathcal{CH}(P)$ of P, and assume that $P \cup \{o\}$ is in general position, i.e. no three points lie on a line.

2 An Exact Algorithm

For a simple polygon $Q = v_0, .., v_n$ whose vertices are given in counterclockwise (ccw) order the well-known shoelace formula gives us that $\mathcal{A}(Q) = \frac{1}{2} \sum_{i=0}^{n} \mathcal{A}'(\overrightarrow{v_i v_{i+1 \bmod n}})$, where $\mathcal{A}'(\overrightarrow{pq}) = \det\left(\begin{smallmatrix} p_x & q_x \\ p_y & q_y \end{smallmatrix}\right)$ denotes the area of the triangle defined by the origin and the directed line segment from p to q. See Fig. 1 for an illustration.

Let E_1, E_2, and E_3 be random indicator variables corresponding to the edges of Δ in ccw order. We then have $\mathbb{E}[\mathcal{A}(\Delta)] =$

$$\mathbb{E}\left[\sum_{i=1}^{3} \mathcal{A}'(E_i)\right] = \sum_{i=1}^{3} \mathbb{E}[\mathcal{A}'(E_i)] = \sum_{i=1}^{3} \sum_{a} a\mathbb{P}[\mathcal{A}'(E_i) = a].$$

*Département d'Informatique, Université Libre de Bruxelles, fisikop@gmail.com

§MADALGO, Aarhus University, [f.staals|constant]@cs.au.dk

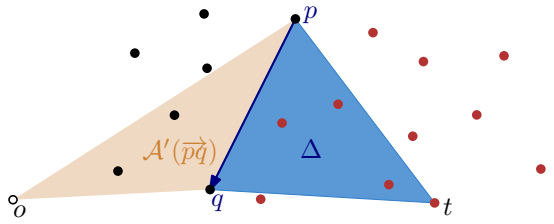

Figure 1: $\mathcal{A}(\Delta)$ is the sum of three "signed" areas, one of which is shown in orange. The number of red points is n_{pq}.

We now observe that all areas a are realized by an ordered pair of points (p, q), and thus $\sum_{i=1}^{m} \sum_{a} a\mathbb{P}[\mathcal{A}'(E_i) = a] =$

$$\sum_{i=1}^{3} \sum_{p,q \in P} \mathcal{A}'(\overrightarrow{pq}) \mathbb{P}[E_i = \overrightarrow{pq}] = \sum_{p,q \in P} \mathcal{A}'(\overrightarrow{pq}) \sum_{i=1}^{3} \mathbb{P}[E_i = \overrightarrow{pq}].$$

An edge $\overrightarrow{pq}$ cannot be both the i^{th} and the j^{th} edge of Δ (for $i \neq j$), and thus, $\sum_{i=1}^{3} \mathbb{P}[E_i = \overrightarrow{pq}]$ equals the probability that $\overrightarrow{pq}$ is a ccw edge in Δ. For $\overrightarrow{pq}$ to be a ccw edge in Δ, the remaining vertex t of Δ should lie to the left of (the oriented line containing) $\overrightarrow{pq}$, and thus $\mathbb{P}[\overrightarrow{pq}$ is a ccw edge in $\Delta] = n_{pq} / \binom{n}{3}$, where n_{pq} is the number of points to (the oriented line containing) $\overrightarrow{pq}$. This is illustrated in Fig. 1. Thus, we have

$$\begin{aligned} \mathbb{E}[\mathcal{A}(\Delta)] &= \sum_{p,q \in P} \mathcal{A}'(\overrightarrow{pq}) \mathbb{P}[\overrightarrow{pq} \text{ is a ccw edge in } \Delta] \\ &= \frac{1}{\binom{n}{3}} \sum_{p,q \in P} \mathcal{A}'(\overrightarrow{pq}) n_{pq}. \end{aligned} \tag{1}$$

As $\mathcal{A}'(\overrightarrow{pq})$ can be computed in $O(1)$ time for every pair p, q, all we need to do is compute all values n_{pq}. We can easily do this in $O(n^2 \log n)$ time, by fixing each point p and sorting the remaining points around p. We conclude:

Theorem 1 *We can compute* $\mathbb{E}[A(\Delta)]$ *in* $O(n^2 \log n)$ *time.*

This approach directly extends to computing $\mathbb{E}[\mathcal{A}(\mathcal{CH}(S))]$ of a randomly selected subset $S \subseteq P$ of size k.

3 A $(1 \pm \varepsilon)$-Approximation

We describe a $(1 \pm \varepsilon)$-approximation algorithm for evaluating Eq. 1, and thus for computing $\mathbb{E}[\mathcal{A}(\Delta)]$. The basic idea is to decompose the $\binom{n}{2}$ pairs of points into few pairs of sets $\{a\}, B$, such that all points $b \in B$ have roughly the same triangle area $\mathcal{A}'(\overrightarrow{ab})$, and to approximate the sum of the n_{ab} values for $b \in B$.

3.1 Approximating the Areas

A well-separated pair decomposition (WSPD), with separation $s = 4/\delta$, of P is a partition of the $\binom{n}{2}$ pairs of points into $m = O(s^2 n \log n)$ pairs of *well-separated* sets $(\{a_i\}, B_i)$, i.e. if B_i fits into a disk $\mathcal{D}(B_i)$ of radius r, the distance between a_i and any point $b \in B_i$ is at least $(s+1)r$ [2]. It follows that for any two points $p, q \in \mathcal{D}(B_i)$, the distance $\|a_i p\|$ is a $(1 \pm \delta)$-approximation of the distance $\|a_i q\|$.

Assume w.l.o.g. that the distance from the origin o to any point in P is at least the diameter d of P. It then follows that for any set B_i, the pair $(\{o\}, B_i)$ is well-separated.

Lemma 2 *Let $(\{a\}, B)$ be a well-separated pair with separation $s = 4/\delta$, where $\delta = \frac{\varepsilon c^2}{40 d^2}$, c is the distance between the closest pair of points in P, and d is the distance between the furthest pair of points in P, and finally let $A_{aB} = \mathcal{A}'(ap)$ for the point $p \in \mathcal{D}(B)$ furthest from the line containing $\overline{ao}$. For every $b \in B$, A_{aB} is a $(1+\varepsilon)$-approximation of $\mathcal{A}'(\overrightarrow{ab})$.*

Lemma 3 *We can compute an oracle that gives an $(1+\varepsilon)$-approximation of $\mathcal{A}'(pq)$, for any $p, q \in P$, in $O(1)$ time, using $O((\rho^4/\varepsilon^2) n \log n)$ preprocessing time, where ρ is the ratio between the furthest and the closest pair of points in P.*

3.2 Approximating the Number of Points

Fix a point $a \in P$ and a subset $B \subseteq P$ of size $z \geq 2$. We present a $(1 \pm \varepsilon)$-approximation for $F^*_a(B) = F^*(B) = \sum_{b \in B} n_b$, where $n_b = n_{ab}$. Our algorithm will compute $F(B) = (z/|B'|) \sum_{b \in B'} n'_b$, where $B' \subseteq B$ is a sample of the points in B, and n'_b is a $(1 \pm \delta)$-approximation of n_b. Let $E = |F^*(B) - F(B)|$ denote the error in our approximation.

Given a line(segment) s, we denote the half planes bounded by the line containing s by s^- and s^+. Let $\mathcal{H} = \{\overline{ab}^+ \mid b \in B\}$ denote the set of half planes defined by a and B. For a given point $p \in P$, let $R_p = \{h \mid h \in \mathcal{H} \wedge h \ni p\}$ denote half planes containing p, and let m_p denote the number of such half planes. We are thus interested in computing $F^*(\mathcal{H}) = \sum_{h \in \mathcal{H}} n_h = \sum_{p \in P} m_p = G^*_{\mathcal{H}}(P)$. To this end, our algorithm distinguishes two cases, depending on z.

Case $\mathcal{H}$ is small. When $z \leq t$, for some, to be determined t, we simply query each plane. Using a $(1 \pm \varepsilon)$-approximate half-plane counting algorithm [1] we immediately get $E \leq \sum_{h \in H} \varepsilon n_p = \varepsilon F^*(\mathcal{H})$.

Case $\mathcal{H}$ is large. When $z > t$ we take a (uniformly drawn) random sample H of the half-planes, and query only the half-planes in H. More precisely, we compute $F(\mathcal{H}) = \overline{F}(H) = (z/|H|) \sum_{h \in H} n'_h$, where n'_h denotes a $(1 \pm \delta)$-approximation of the number of points from P on half-plane h. If we take a sample of size $O(r^2 \log r)$, then whp. H is an $(1/r)$-approximation for the range space $(\mathcal{H}, \mathcal{R})$, where $\mathcal{R} = \{R_p \mid p \in P\}$ [3]. That is, for all ranges $R \in \mathcal{R}$ we have that

$$\left| \frac{|R|}{|\mathcal{H}|} - \frac{|R \cap H|}{|H|} \right| \leq (1/r).$$

This allows us to show that the absolute error E is at most $nz/r + nz\delta$. We now choose $(1/r) = \delta = (z\varepsilon)/8n$, which gives us $E \leq \varepsilon z^2/4$. By ordering the points defining the planes in $\mathcal{H}$ appropriately, we get $F^*(\mathcal{H}) \geq z(z-1)/2 \geq z^2/4$. Thus, $F(\mathcal{H})$ is a $(1 \pm \varepsilon)$-approximation for F^*.

Running time. We choose the threshold t to minimize the running time. If $\mathcal{H}$ is small the running time to handle the pair $(\{a\}, B)$ is $O(z \log n)$. If $\mathcal{H}$ is large the running time is $O(r^2 \log r \log n) = O(\frac{n^2}{z^2 \varepsilon^2} \log^2 n)$. These two quantities balance out for $t = z = (n/\varepsilon)^{2/3} \log^{1/3} n$. We conclude:

Lemma 4 *After $O(n \log n)$ expected time preprocessing, we can whp. compute a $(1 \pm \varepsilon)$ approximation of $F^*_a(B)$, for any $\{a\} \cup B \subseteq P$, in $O((n/\varepsilon)^{2/3} \log^{1/3} n)$ expected time.*

3.3 Combining the Approximations

Straightforward calculations show that if we combine the results from Lemmas 3 and 4, choosing both approximation errors to be $\varepsilon/3$, we get a $(1 \pm \varepsilon)$-approximation.

Theorem 5 *Whp. we can compute a $(1 \pm \varepsilon)$-approximation of $\mathbb{E}[\mathcal{A}(\Delta)]$ in $O(\frac{1}{\varepsilon^{8/3}} \rho^4 n^{5/3} \log^{4/3} n)$ expected time.*

4 Future Work

We would like to improve, or remove, the dependency on ρ in our approximation algorithm. A possible approach to do so would be to replace the WSPD by a different decomposition of the pairs of points that allows a better approximation of the triangle areas. We conjecture that computing $\mathbb{E}[\mathcal{A}(\Delta)]$ exactly is 3SUM-hard. Proving this is another avenue for future work. Finally, we would like to investigate a $(1 \pm \varepsilon)$-approximation algorithm for the general problem of computing $\mathbb{E}[\mathcal{A}(\mathcal{CH}(S))]$ for a fixed size sample S.

References

[1] P. Afshani and T. M. Chan. On Approximate Range Counting and Depth. *Discrete & Comp. Geom.*, 42(1):3–21, 2009.

[2] P. B. Callahan and S. R. Kosaraju. A decomposition of multidimensional point sets with applications to k-nearest-neighbors and n-body potential fields. *J. ACM*, 42(1):67–90, Jan. 1995.

[3] D. Haussler and E. Welzl. ε-nets and simplex range queries. *Discrete & Comp. Geom.*, 2(2):127–151, 1987.

[4] M. A. Mouchet, S. Villéger, N. W. Mason, and D. Mouillot. Functional diversity measures: an overview of their redundancy and their ability to discriminate community assembly rules. *Functional Ecology*, 24(4):867–876, 2010.